7906

यह पुस्तक वितरण [illegible]
NOT TO BE ISSUED

सन्दर्भ ग्रन्थ
REFERENCE BOOK

स्थानक नगरीकरण १९८४-१९८५

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

390.6
Vol.31

REVIEWS OF MODERN PHYSICS

CHECKED 1973
Initial ...

BOARD OF EDITORS

S. A. GOUDSMIT, *Managing Editor* E. U. CONDON, *Editor*

Associate Editors

CHARLES KITTEL
G. C. WICK
D. R. INGLIS
PHILIP MORRISON
JOHN BARDEEN
ROBERT HOFSTADTER

VOLUME 31
1959

पुस्तकालय
पुस्तक संख्या 530.5
AM3R
तिथि V.31
गुरुकुल कांगड़ी विश्वविद्यालय

Published for the

AMERICAN PHYSICAL SOCIETY

by the

AMERICAN INSTITUTE OF PHYSICS
Incorporated

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

REVIEWS OF MODERN PHYSICS

Reviews of Modern Physics is published quarterly to provide those interested in developments in physics with comprehensive and timely discussions of the problems which are of especial interest and importance.

It is hoped that the studies in this journal will give to the specialist the benefit of the viewpoint of other specialists; to the student, a background of critical knowledge and an adequate bibliography of source material; to those whose time is absorbed in the teaching of physics, a stimulating account of the forward movement of their science.

Published by the American Institute of Physics for the

AMERICAN PHYSICAL SOCIETY

OFFICERS

President

G. E. UHLENBECK
University of Michigan
Ann Arbor, Michigan

Vice President

V. F. WEISSKOPF
Massachusetts Institute of Technology
Cambridge 39, Massachusetts

Secretary

K. K. DARROW
Bell Telephone Laboratories
(*retired*)
Office:
Columbia University
New York 27, New York

Treasurer

S. L. QUIMBY
Columbia University
New York 27, New York

Local Secretary for the Pacific Coast

WILLIAM A. NIERENBERG
University of California
Berkeley, California

Managing Editor

S. A. GOUDSMIT
Brookhaven National Laboratory
Upton, Long Island
New York

BOARD OF EDITORS

Managing Editor
S. A. GOUDSMIT

Editor
E. U. CONDON

Associate Editors

CHARLES KITTEL
G. C. WICK
D. R. INGLIS
PHILIP MORRISON
JOHN BARDEEN
ROBERT HOFSTADTER

Manuscripts for publication should be submitted to E. U. Condon, Department of Physics, Washington University, St. Louis 5, Missouri.

Proof and all correspondence concerning papers in the process of publication should be addressed to the Publication Manager, American Institute of Physics, 335 East 45 Street, New York 17, New York.

Subscription Price

	U. S. and Canada	Elsewhere
To members of the American Physical Society	$4.50	$4.90
To all others	6.00	6.40

Back Numbers

Yearly rate when complete year is available: **$12.00**
Single copies: **$3.50** each.
Double number: **$7.00**

Subscriptions, renewals, and orders for back numbers should be addressed to the American Institute of Physics, 335 East 45 Street, New York 17, New York.

Changes of address should be addressed to the American Institute of Physics.

Reviews of Modern Physics is published quarterly in January, April, July, and October at Prince and Lemon Streets, Lancaster, Pennsylvania.

Second-class postage paid at Lancaster, Pa.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

Contents

JANUARY, 1959

BIOPHYSICAL SCIENCE

A STUDY PROGRAM

Acknowledgments ii
Biophysical Science—A Study Program 1
Selected General Bibliography 2
Molecular Biology and the Physical Basis of Life Processes . . . . . Francis O. Schmitt 5
Cellular Dynamics . . . . . Paul Weiss 11
Cellular Responses . . . . . William Bloom 21
Chemical Characterization of Proteins, Carbohydrates, and Lipids . . . . . J. L. Oncley 30
Molecular Configuration of Synthetic and Biological Polymers . . . . . Alexander Rich 50
Physicochemical Characterization of Macromolecules . . . . . Paul Doty 61
Polyelectrolytes . . . . . Stuart A. Rice 69
Proteins and Their Interactions . . . . . David F. Waugh 84
Three-Dimensional Structure of Globular Proteins . . . . . John C. Kendrew 94
The Hydrogen Bond . . . . . Leslie E. Orgel 100
Forces between Macromolecules . . . . . Walter H. Stockmayer 103
Configurations of Biologically Important Macromolecules in Solution . . . . . Paul Doty 107
Application of Infrared Spectroscopy to Biological Problems . . . . . G. B. B. M. Sutherland 118
Concentrated Macromolecular Solutions . . . . . Bruno H. Zimm 123
Rheology of Macromolecular Systems . . . . . John D. Ferry 130
Respiratory-Energy Transformation . . . . . Albert L. Lehninger 136
Energy Reception and Transfer in Photosynthesis . . . . . Melvin Calvin 147
Free Radicals in Photosynthetic Systems . . . . . Melvin Calvin 157
Relation between Exciton Bands and Conduction Bands in Molecular Lamellar Systems . Michael Kasha 162
General Patterns of Biochemical Synthesis . . . . . Richard Brooke Roberts 170
Mechanisms of Enzyme Action . . . . . Robert A. Alberty 177
Protein Structure and Enzyme Action . . . . . Hans Neurath 185
Molecular Structure of the Nucleic Acids . . . . . Alexander Rich 191
Biosynthesis of Nucleic Acids . . . . . Arthur Kornberg 200
Approaches to the Biosynthesis of Proteins . . . . . Alton Meister 210
Coding and Information Theory . . . . . Peter Elias 221
Genetics and the Role of Nucleic Acids . . . . . Cyrus Levinthal 227
Replication of Nucleic Acids . . . . . Robley C. Williams 233
Genetic Fine-Structure Analysis . . . . . E. S. Lennox 242
Coding Aspects of Protein Synthesis . . . . . Cyrus Levinthal 249
Induced Enzyme Synthesis . . . . . David S. Hogness 256

APRIL, 1959

Biophyscial Science. A Study Program i
General Features of Radiobiological Actions . . . . . Raymond E. Zirkle 269
Radiation Inactivation of Enzymes, Nucleic Acids, and Phage Particles . . . . . Ernest Pollard 273
Some Aspects of Cellular Radiobiology . . . . . T. H. Wood 282
Effects of Radiations on Populations of Cells and Multicellular Organisms . . . . . Cornelius A. Tobias 289
Fine Structure of Cell Nucleus, Chromosomes, Nucleoli, and Membrane . . . . . H. Stanley Bennett 297
Fine Structure of Cytoplasm: The Organization of Membranous Layers . . . . . Fritiof S. Sjöstrand 301
Fine Structure of Biological Lamellar Systems . . . . . H. Fernández-Morán 349
Fine Structure of Lamellar Systems as Illustrated by Chloroplasts . . . . . Alan J. Hodge 331
Membrane Transport . . . . . Robert W. Berliner 342
Interaction Properties of Elongate Protein Macromolecules with Particular Reference to Collagen (Tropocollagen) . . . . . Francis O. Schmitt 349
Molecular Biology of Mineralized Tissues with Particular Reference to Bone . . . . Melvin J. Glimcher 359
Structure of Muscle Cells . . . . . H. Stanley Bennett 394
Motility of Cilia and the Mechanism of Mitosis . . . . . Shinya Inoué 402
Fibrous Proteins of Muscle . . . . . Alan J. Hodge 409
Mechanisms of Muscle Contraction . . . . . Manuel F. Morales 426

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

Quantitative Studies on Mammalian Cells *in Vitro* . . . . . T. T. Puck 433
Interactions between Cells . . . . . Paul Weiss 449
Molecular Organization of the Nerve Fiber . . . . . Francis O. Schmitt 455
Nature of the Nerve Impulse . . . . . Bernard Katz 466
Recovery Processes and Metabolism of Nerve . . . . . C. M. Connelly 4[illegible]
Sensory Performance of Organisms . . . . . Walter A. Rosenblith [illegible]
Biological Transducers and Coding . . . . . Otto H. Schmitt 492
Initiation of Nerve Impulses in Receptor and Central Neurons . . . . . Theodore Holmes Bullock 504
Receptor Mechanisms and the Integration of Sensory Information in the Eye . . . . . H. K. Hartline 515
Mechanisms of Synaptic Transmission . . . . . Bernard Katz 524
Some Quantifiable Aspects of the Electrical Activity of the Nervous System (with Emphasis upon Responses to Sensory Stimuli) . . . . . Walter A. Rosenblith 537
Imitation of Pattern Recognition and Trial-and-Error Learning in a Conditional Probability Computer . . . . . Albert M. Uttley 546
Chemical Specificity in Biological Systems . . . . . Walter Kauzmann 549
Blood Coagulation—A Study in Homeostasis . . . . . David F. Waugh 557
Hormone Regulation . . . . . DeWitt Stetten, Jr. 563

JULY, 1959

Structure of Spectral Lines from Plasmas . . . . . Henry Margenau and Marvin Lewis 569
Electronic Band Structure of Solids by X-Ray Spectroscopy . . . . . L. G. Parratt 616
Properties of Rutile (Titanium Dioxide) . . . . . F. A. Grant 646
Comparison of the Debye θ Determined from Elastic Constants and Calorimetry . . . . . G. A. Alers and J. R. Neighbours 675
Masers . . . . . J. Weber 681
Polarization Measurements on Nuclear Gamma Rays . . . . Lawrence W. Fagg and Stanley S. Hanna 711
Electron and Positron Polarization . . . . . Lorne A. Page 759
Invited Papers from the Conference on Weak Interactions Held at Gatlinburg, Tennessee, October 27–29, 1958 782
Parity Experiments in Beta Decays . . . . . C. S. Wu 783
Determination of the Beta-Decay Interaction from Electron-Neutrino Angular Correlation Experiments . . . . . James S. Allen 791
Mu Capture, Beta Decay, and Pi-Meson Decay . . . . . M. L. Goldberger 797
Theory of Muon Capture . . . . . H. Pri[illegible]koff 802
Strange Particle Decay Processes and the Fermi Interaction . . . . . R. H. Dalitz 823
Status of Weak Interactions . . . . . Murray Gell-Mann 834
Errata: 839

Comparative Study of Potential Energy Functions for Diatomic Molecules . . . . . Yatendra Pal Varshni 839
Experimental Results on Charge-Changing Collisions of Hydrogen and Helium Atoms and Ions at Kinetic Energies above 0.2 kev . . . . . Samuel K. Allison 839
Spin Waves . . . . . J. Van Kranendonk and J. H. Van Vleck 839
Dynamics of a Lattice Universe by the Schwarzschild-Cell Method . Richard W. Lindquist and John A. Wheeler 839
Theoretical Results on Orbital Capture . . . . . H. Brysk and M. E. Rose 840
Pair Production and Bremsstrahlung in the Field of Free and Bound Electrons . . . J. Joseph and F. Rohrlich 840

OCTOBER, 1959

Internal Rotation and Microwave Spectroscopy . . . . . Chun C. Lin and Jerome D. Swalen 8[illegible]
Foundations of the Optical Model for Nuclei and Direct Interaction . . . . . G. E. Brown 8[illegible]
Bremsstrahlung Cross-Section Formulas and Related Data . . . . . H. W. Koch and J. W. Motz 9[illegible]
Interaction of Normal Modes with Electron Traps . . . . . Jordan J. Markham 9[illegible]
Energy Transport in Organic Phosphors . . . . . S. C. Ganguly and N. K. Chaudhury 9[illegible]
Irreversible Thermodynamics of Nonlinear Processes and Noise in Driven Systems . . . . . William Bernard and Herbert B. Callen 10[illegible]
Review of High-Temperature Rotating-Plasma Experiments . . . . . John M. Wilcox 10[illegible]
Ultrasonic Absorption Due to Chemical Relaxation in Electrolytes . . . . . G. S. Verma 10[illegible]
Anisotropic Light Scattering of Streaming Suspensions and Solutions . . . . . Wilfried Heller 1[illegible]
Volume Table of Contents . . . . . [illegible]

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

112457

112457

REVIEWS OF MODERN PHYSICS

VOLUME 31, NUMBER 1 JANUARY, 1959

BIOPHYSICAL SCIENCE

A Study Program

Planned and Edited by

J. L. Oncley, *Editor-in-Chief*

F. O. Schmitt

R. C. Williams

M. D. Rosenberg

R. H. Bolt

for the

Biophysics and Biophysical Chemistry Study Section

of the National Institutes of Health

Public Health Service

United States Department of Health, Education, and Welfare

Published in the

Reviews of Modern Physics, January and April, 1959

and by

John Wiley and Sons, Inc., New York

Copyright © 1959 by the American Physical Society

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

Acknowledgments

The Study Program in Biophysical Science, Boulder, Colorado, 20 July–16 August 1958, was organized and conducted by the Biophysics and Biophysical Chemistry Study Section of the National Institutes of Health, Public Health Service, United States Department of Health, Education, and Welfare, under research grant RG-5048 from that agency. About 120 senior research contributors and selected younger scientists took part in this program, which was held in an environment conducive to informal interchange of ideas.

The general objective of the Study Program was to aid and encourage the further blending of concepts and methods of physical science with those of life science in the investigation of biological problems. The Study Section conceived this activity as an experiment in the exchange of information and ideas among representatives of different branches of science. The core of the Study Program was a carefully integrated series of about 60 lectures, constituting compact summaries of certain key problems and critical evaluations of recent advances. The lecture material provided a framework for other activities such as planned workshop sessions, spontaneous discussion groups, and library study.

The results of the experiment reinforced the view of the Study Section that the same lecture material, appropriately edited and published, would also provide a timely research guide and a useful base on which to build new courses and seminars in biophysical science. Thus, the values of this effort would be made available to a much broader audience than could be accommodated in the informal atmosphere desired for the Study Program.

The method of publication reflects the objectives of the Study Program. Presentation in the *Reviews of Modern Physics* brings the material before the community of physicists, who may find interest and stimulation here. In this respect, the publication is a sequel to "Borderland Problems in Biology and Physics" by John R. Loofbourow, which appeared in the *Reviews of Modern Physics* during 1940. Simultaneous publication in book form by John Wiley and Sons, Inc., who have prepared the illustration drawings, brings the material to the attention of scientists in all fields and countries. Through the combined publication procedure, the results of the Study Program are being disseminated promptly, to a wide audience, at the lowest practicable cost.

The entire Study Section devoted long hours and much thought to the formulation of the Study Program. Encouragement and help were given by former members T. Hill and G. B. B. M. Sutherland. An invaluable aid in formulating the program was the experience gained from the "Conference on Certain Fundamental Aspects of Biophysical Science," conducted by the Study Section in January 1958 at Bethesda, Maryland, and planned by a committee: P. H. Abelson, R. C. Williams, and H. Neurath, chairman.

The detailed planning and final decisions regarding subject matter of the summer Study Program were mainly the responsibility of a committee: R. H. Bo[illegible] J. L. Oncley, M. D. Rosenberg, R. C. Williams, F. O. Schmitt, chairman. Valuable help was pro[illegible] also by H. Fernández-Morán, A. Rich, C. Levin[illegible] and D. F. Waugh.

Able technical editorial assistance was given considerable number of the Study Program participan[illegible] especially R. L. Baldwin, M. S. Blois, Jr., A. S. Brin[illegible] L. S. Frishkopf, M. J. Glimcher, M. H. Goldstein, Jr., P. R. Gross, R. G. Hart, R. I. Henkin, M. Kasha, G. Kegeles, M. S. Meselson, H. H. Pattee, Jr., F. W. Putnam, S. Shulman, I. W. Sizer, and I. Tincoco, Jr. In the revision of manuscripts, special assistance was given by M. V. Edds, Jr., and W. Bloom.

This publication would not have been possible without the devoted and expert assistance of Jean Doty. She has carried much of the burden of editing, proofreading, and coordination of manuscript preparation. We are grateful to G. W. Kolodny and W. Scheider for collecting bibliographic material and checking references; to G. F. Gove for undertaking the difficult job of indexing; to M. E. Platin for extensive secretarial and administrative assistance.

The combined publication was made possible by the active interest and cooperation of the editor of the *Reviews of Modern Physics*, E. U. Condon, who also was a participant in the Study Program. We are indebted to Ruth Bryans, Publication Manager of the American Institute of Physics, and staff; and to many other friends in scientific institutions and publishing houses who have advised and encouraged us in this undertaking. We are indebted also to the many authors, societies, and publishers for the use of figures, tables, and other illustrative material; specific acknowledgments are given where they appear in the papers.

The Study Program participants, in addition to those named elsewhere in these acknowledgments or listed as authors of papers, were: B. Altshuler, W. T. Astbury, L. G. Augenstine, J. J. Baruch, E. D. Becker, S. A. Bernhard, H. B. Callen, F. D. Carlson, W. R. Carroll, N. R. Davidson, D. R. Davies, E. A. Edelsack, M. Eden, A. V. Engström, D. J. Fluke, F. L. Friedman, H. E. von Gierke, J. Z. Hearon, D. F. Hornig, S. H. Hutner, H. S. Kaplan, N. O. Kaplan, J. Kraut, U. Liddel, L. Lorand, E. F. MacNichol, Jr., P. Morrison, L. J. Mullins, J. R. Platt, R. Podolsky, E. Ribi, S. J. Singer, R. L. Sinsheimer, H. W. Smith, M. R. Stetten, L. Szilard, E. W. Taylor, T. E. Thompson, C. H. Townes, J. Towsend, B. L. Vallee, A. C. Young.

To all participants, we express our gratitude for critical discussions of the scientific material and cogent suggestions regarding its presentation. It is they as a group who have vitalized this account of horizons in biophysical science.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

Biophysical Science—A Study Program

Physical and chemical approaches to problems in biology have become increasingly productive in recent years. Major advances in the understanding of life processes have been made through research in such specialties as biophysical chemistry, molecular biology, biophysics, and electrophysiology. Continuing progress will require an ever more perceptive study of the interactions of matter, energy, and information in biological systems.

This publication grew out of a special activity, designed to aid and stimulate the further blending of the concepts and methods of physics and chemistry with those of the life sciences in the study of biological problems. The papers in this volume were presented in the Study Program in Biophysical Science, held in Boulder, Colorado, during the summer of 1958.

A living system is a self-perpetuating combination of atoms, organized in a highly specific manner and interacting with its environment in such a way as to metabolize, reproduce, grow, and adapt. The Study Program was designed to present biological problems as "viewed through physical spectacles and investigated by physical ideas and methods."* Biological processes can be examined, for example, in terms of atomic structures, energy levels, and binding forces. The engineer might prefer to specify the design and operation in terms of components and "black boxes." The chemist discusses molecular configurations and kinetic and thermodynamic details of biochemical reactions. These and other modes for describing biological processes are found in this volume.

The Study Program included a number of "case studies" in which topics were chosen to point up the need and value of physical approaches at all levels of biological organization. Particular emphasis centered upon muscle and nerve, where considerable information is available at many of these levels. The series of papers related to nerve, for example, encompasses the molecular organization of the nerve fiber, the nature of the nerve impulse, nerve metabolism, sensory performance of organisms, receptor mechanisms, and integrated responses of higher neural centers. Another series deals with genetics, replication, and synthesis of nucleic-acid and protein macromolecules. Here again, the topics range from descriptions of the biological processes of replication, through physical and biochemical details of relevant structures, to mathematical considerations in the coding of information.

The subjects were selected with emphasis on fundamental concepts, to broaden the base of common language and understanding among a heterogeneous group of biologists, physicists, and chemists. This emphasis restricted the range of topics that could be presented within the limited duration of the Study Program. Also, the interests of those who organized the program are unavoidably reflected in its contents.

This particular choice of topics is not intended to define or delimit the field of biophysics, but, rather, to illustrate the power and success of the integrated physical-biological approach. Future conferences and publications of this kind will undoubtedly emphasize alternative subjects such as, for example, those at higher levels of biological complexity.

The arrangement of the material presented in the Study Program was chosen to provide a logical unfolding of the selected subjects, along with a continuous injection of background material. The initial chapters give many of the fundamental concepts, whereas the later chapters make use of developments introduced in the intervening chapters. Since references may be needed to supplement the necessarily abridged discussions in the papers, a Selected General Bibliography is given on pages 2 to 4. The entries are divided into 10 subject categories, together with a designation of those papers that are most closely related to each category.

Biophysics and Biophysical Chemistry Study Section

P. H. Abelson
J. W. Beams
R. H. Bolt
J. D. Ferry
I. Fuhr
D. E. Goldman
I. Gray
J. D. Hardy
H. K. Hartline
J. G. Kirkwood
H. Neurath
J. L. Oncley
M. D. Rosenberg
H. A. Sober
R. C. Williams
F. O. Schmitt, *Chairman*

October, 1958
Cambridge, Massachusetts

* A. V. Hill, "Why Biophysics?" *Lectures on the Scientific Basis of Medicine* (Athlone Press, London, 1956), Vol. IV, p. 1. Reprinted in Science 124, 1233 (1956).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

Selected General Bibliography

Cellular Biology (related papers: 1–3, 47, 48)

1. Frey-Wyssling, A., *Submicroscopic Morphology of Protoplasm* (Elsevier Publishing Company, New York, 1953).
2. Giese, A. C., *Cell Physiology* (Saunders Company, Philadelphia, 1957).
3. Conference on Tissue Fine Structure, Biophys. Biochem. Cytol. **2**, Suppl., 1–448 (1956).
4. McElroy, W. D., and Glass, B., editors, *The Chemical Basis of Heredity* (The Johns Hopkins Press, Baltimore, Maryland, 1957).
5. Mellors, R. C., *Analytical Cytology* (McGraw-Hill Book Company, Inc., New York, 1955).
6. Schmidt, W. J., *Die Doppelbrechung von Karyoplasma, Zytoplasma, und Metaplasma* (Verlage Borntraeger, Berlin, 1937).
7. Schmitt, F. O., Doty, P., Hall, C. E., Williams, R. C., and Weiss, P. A., Symposium on Biomolecular Organization and Life-Processes, Proc. Natl. Acad. Sci. U. S. **42**, 789–830 (1956).
8. Willier, B. H., Weiss, P. A., and Hamburger, V., editors, *Analysis of Development* (Saunders Company, Philadelphia, 1955).
9. Wilson, E. B., *The Cell in Development and Heredity* (The Macmillan Company, New York, 1925).

Physical and Chemical Characteristics of Macromolecules (related papers: 4–15, 23).

1. Chargaff, E., and Davidson, J. N., editors, *The Nucleic Acids* (Academic Press, Inc., New York, 1955).
2. Davidson, J. N., *The Biochemistry of the Nucleic Acids* (Methuen Company, London; and John Wiley and Sons, Inc., New York, 1957), third edition.
3. Edsall, J. T., and Wyman, J., *Biophysical Chemistry* (Academic Press, Inc., New York, 1958), Vol. I.
4. Eirich, F. R., editor, *Rheology: Theory and Applications* (Academic Press, Inc., New York, 1956–1958).
5. Flory, P. J., *Principles of Polymer Chemistry* (Cornell University Press, Ithaca, New York, 1953).
6. Fruton, J. S., and Simmonds, S., *General Biochemistry* (John Wiley and Sons, Inc., New York, 1958), second edition.
7. Katchalsky, A., "Problems in the physical chemistry of polyelectrolytes," J. Polymer Sci. **12**, 159–184 (1954).
8. Mark, H., and Tobolsky, A. V., *Physical Chemistry of High Polymeric Systems* (Interscience Publishers, Inc., New York, 1950), second edition.
9. Neuberger, A., editor, *Symposium on Protein Structure* (John Wiley and Sons, Inc., New York; and Methuen Company, London, 1958).
10. Neurath, H., and Bailey, K., editors, *The Proteins: Chemistry, Biological Activity and Methods* (Academic Press, Inc., New York, 1953).
11. Springall, H. D., *The Structural Chemistry of Proteins* (Butterworths Scientific Publications, London, 1954).

Energy Transfer and Biochemical Synthesis (related papers: 16–22, 24, 25)

1. Arnold, W., and Sherwood, H. K., "Are chloroplasts semiconductors?" Proc. Natl. Acad. Sci. U. S. **43**, 105–114 (1957).
2. Bassham, J. A., and Calvin, M. C., *The Path of Carbon in Photosynthesis* (Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1957).
3. Chance, B., and Williams, G. R., "The respiratory chain and oxidative phosphorylation," Advances in Enzymol. **17**, 65–134 (1956).
4. Gutfreund, H., "The nature of entropy and its role in biochemical processes," Advances in Enzymol. **11**, 1–33 (1951).
5. Karreman, G., and Steele, R. H., "On the possibility of long distance energy transfer by resonance in biology," Biochim. et Biophys. Acta **25**, 280–291 (1957).
6. Krebs, H. A., and Kornberg, A. L., *Energy Transformations in Living Matter* (Springer-Verlag, Berlin, 1957).
7. Lehninger, A. L., "Oxidative phosphorylation" in *The Harvey Lectures 1953–1954*, The Harvey Society of New York, Series XLIX (Academic Press, Inc., New York, 1955), pp. 176–215.
8. Lehninger, A. L., Wadkins, C. L., Cooper, C., Devlin, T. M., and Gamble, J. L., Jr., "Oxidative phosphorylation," Science **128**, 450–456 (1958).
9. Reid, C., *Excited States in Chemistry and Biology* (Academic Press, Inc., New York, 1957).
10. Sogo, P. B., and Tolbert, B. M., "Nuclear and electron paramagnetic resonance and its application to biology," in *Physics*, J. H. Lawrence and C. A. Tobias, editors (Academic Press, Inc., New York, 1957), Vol. V, pp. 1–35.

Genetics and Replication of Proteins and Nucleic Acids (related papers: 26–31)

1. Cherry, C., editor, *Information Theory* (Butterworths Scientific Publications, London, 1956).
2. Cohn, M., and Monod, J., "Specific inhibition and induction of enzyme biosynthesis," in *Adaptation in Microorganisms*, R. Davies and E. F. Gale, editors (Cambridge University Press, Cambridge, England, 1953), pp. 132–149.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

3. "Viruses," Cold Spring Harbor Symposia Quant. Biol. **18**, (1953).
4. Dulbecco, R., "Interaction of viruses and animal cells—a study of facts and interpretations," Physiol. Rev. **35**, 301–335 (1955).
5. Feinstein, A., *Foundations of Information Theory* (McGraw-Hill Book Company, Inc., New York, 1958).
6. Hershey, A. D., "Bacteriophages as genetic and biochemical systems," Advances in Virus Research **4**, 25–62 (1957).
7. Jacob, F., *Les Bactéries Lysogènes et la Notion de Provirus* (Masson et Cie., Paris, 1954).
8. McElroy, W. D., and Glass, B., editors, *A Symposium on the Chemical Basis of Heredity* (The Johns Hopkins Press, Baltimore, Maryland, 1957).
9. Pontecorvo, G., "Genetic formulation of gene structure and gene action," Advances in Enzymol. **13**, 121–149 (1952).
10. Shannon, C. E., and Weaver, W., *The Mathematical Theory of Communication* (University of Illinois Press, Urbana, Illinois, 1949).
11. Spiegelman, S., and Campbell, A. M., "The Significance of induced enzyme formation," in *Currents in Biochemical Research*, D. E. Green, editor (Interscience Publishers, Inc., New York, 1956), pp. 115–161.
12. Stanier, R. Y., "Enzymatic adaptation in bacteria," Ann. Rev. Microbiol. **5**, 35–56 (1951).

Biological Effects of Radiation (related papers: 32–35)

1. Bacq, Z. M., and Alexander, P., *Fundamentals of Radiobiology* (Academic Press, Inc., New York, 1955).
2. Hollaender, A., editor, *Radiation Biology*. Vol. I: High Energy Radiation (1954); Vol. II: Ultraviolet and Related Radiations (1955); Vol. III: Visible and Near-visible Light (1956) (McGraw-Hill Book Company, Inc., New York, 1954–1956).
3. Lea, D. E., *Actions of Radiations on Living Cells* (Cambridge University Press, New York, 1955), second edition.
4. Nickson, J. J., editor, *Symposium on Radiobiology* (John Wiley and Sons, Inc., New York, 1952).

Molecular Organization and Function (related papers: 36–40)

1. "Active transport and secretion," Symp. Soc. Exptl. Biol. **8**, (1954).
2. Brachet, J., *Biochemical Cytology* (Academic Press, Inc., New York, 1957).
3. Clarke, H. T., editor, *Ion Transport Across Membranes* (Academic Press, Inc., New York, 1954).
4. Conference on Tissue Fine Structure, J. Biophys. Biochem. Cytol. **2**, Suppl. (1956).
5. DeRobertis, E. D. P., Nowinski, W. W., and Saez, F. A., *General Cytology* (Saunders Company, Philadelphia, 1954).
6. Engström, A., and Finean, J. B., *Biological Ultrastructure* (Academic Press, Inc., New York, 1958).
7. Frey-Wyssling, A., *Macromolecules in Cell Structure* (Harvard University Press, Cambridge, 1957).
8. Mellors, R. C., editor, *Analytical Cytology* (McGraw-Hill Book Company, Inc., New York, 1955).
9. "Mitochondria and other cytoplasmic inclusions," Symp. Soc. Exptl. Biol. **10**, (1957).
10. Oster, G., and Pollister, A. W., editors, *Physical Techniques in Biological Research*. Vol. III: Cells and Tissues (Academic Press, Inc., New York, 1956).
11. Schmitt, F. O., "Tissue ultrastructure analysis: polarized light method," in *Medical Physics*, O. Glasser, editor (Yearbook Publishers, Chicago, 1944), pp. 1586–1591.

Connective Tissue and Muscle (related papers: 41–46)

1. "Fibrous proteins and their biological significance," Symp. Soc. Exptl. Biol. **9**, (1955).
2. Gustavson, K. H., *The Chemistry and Reactivity of Collagen* (Academic Press, Inc., New York, 1956).
3. O'Flaherty, F., editor, *The Chemistry and Technology of Leather* (Rheinhold Publishing Corporation, New York, 1956).
4. Peters, R., Hill, A. V., Huxley, A. F., Huxley, H. E., Wilkie, D. R., Bailey, K., Perry, S. V., Needham, D. M., Pantin, C. F. A., Gray, J., Katz, B., *et al.*, "Physiology of voluntary muscle," Brit. Med. Bull. **12**, 161–235 (1956).
5. Randall, J. T., "Observations on the collagen system," J. Intern. Soc. Leather Trades' Chemists **38**, 362–387 (1954).
6. Schmitt, F. O., "Macromolecular interaction patterns in biological systems," Proc. Am. Phil. Soc. **100**, 476–486 (1956).
7. Schmitt, F. O., "Giant molecules in cells and tissues," Scientific American **197**, 204–216 (1957).
8. Stainsby, G., editor, *Recent Advances in Gelatin and Glue Research*, (Pergamon Press, Inc., New York, 1958).
9. Szent-Györgyi, A., *Chemistry of Muscular Contraction* (Academic Press, Inc., New York, 1951), second edition.
10. Tunbridge, R. E., editor, *Connective Tissue: A C.I.O.M.S. Symposium* (Blackwell Scientific Publications, Oxford, 1957).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

Nerve (related papers: 49–51)

1. Fernández-Morán, H., "The submicroscopic structure of nerve fibers," Progr. in Biophys. and Biophys. Chem. **4**, 112–147 (1954).
2. Geren, B. B., "Structural studies of the formation of the myelin sheath in peripheral nerve fibers," in *The Fourteenth Symposium of the Society for the Study of Growth and Development*, D. Rudnick, editor (Princeton University Press, Princeton, New Jersey, 1956), pp. 213–219.
3. Hodgkin, A. L., "Ionic movements and electrical activity in giant nerve fibres," Proc. Roy. Soc. (London) **148B**, 1–37 (1958).
4. Katz, B., "Microphysiology of the neuro-muscular junction. A physiological 'quantum of action' at the myoneural junction," Bull. Johns Hopkins Hosp. **102**, 275–296 (1958).
5. Palay, S. L., "Structure and function in the neuron," in *Progress in Neurobiology and Neurochemistry*, S. R. Korey and J. I. Nurnberger, editors (Paul B. Hoeber, New York, 1956), Vol. I, pp. 64–82.
6. Richter, D., editor, *Metabolism of the Nervous System* (Pergamon Press, New York, 1957).
7. Schmitt, F. O., and Geschwind, N., "The axon surface," Progr. in Biophys. and Biophys. Chem. **8**, 165–215 (1957).
8. Shedlovsky, T., editor, *Electrochemistry in Biology and Medicine* (John Wiley and Sons, Inc., New York, 1955).
9. Waelsch, H., *Ultrastructure and Cellular Chemistry of Neural Tissue* (Paul B. Hoeber, New York, 1957).
10. Weiss, P., *Genetic Neurology* (The University of Chicago Press, Chicago, 1950).

Sensory Reception and Signal Processing (related papers: 52–58)

1. Adrian, E. D., Bremer, F., and Jasper, H. H., editors, *Brain Mechanisms and Consciousness* (Blackwell Scientific Publications, Oxford, 1954).
2. Broadbent, D. E., *Perception and Communication* (Pergamon Press, New York, 1958).
3. Bruner, J. S., "Neural mechanisms in perception," Psychol. Rev. **64**, 340–358 (1957).
4. Cherry, C., editor, *On Human Communication* (John Wiley and Sons, Inc., New York, 1957).
5. Eccles, J. C., *The Physiology of Nerve Cells* (The Johns Hopkins Press, Baltimore, Maryland, 1957).
6. Geldard, F., *The Human Senses* (John Wiley and Sons, Inc., New York, 1953).
7. Granit, R., *Receptors and Sensory Perception* (Yale University Press, New Haven, 1955).
8. Magoun, H. W., *The Waking Brain* (Charles C. Thomas, Springfield, Illinois, 1958).
9. Miller, G. A., "Human memory and the storage of information," Inst. Radio. Engrs. Trans. PGIT **2**, 129–137 (1956).
10. von Neumann, J., *The Computer and the Brain* (Yale University Press, New Haven, 1958).
11. de Vries, Hl., "Physical aspects of the sense organs," Progr. in Biophys. and Biophys. Chem. **6**, 207–264 (1956).
12. Wiener, N., *Nonlinear Problems in Random Theory* (Technology Press, Cambridge; and John Wiley and Sons, Inc., New York, 1958).

Specificity in the Chemical Control of Biological Systems (related papers: 59–61)

1. Blum, H. F., *Time's Arrow and Evolution* (Princeton University Press, Princeton, New Jersey, 1951).
2. Ferguson, J. H., "General properties of blood: the formed elements," pp. 503–520; "Coagulation of blood," pp. 540–556; in *Textbook of Physiology*, J. F. Fulton, editor (Saunders Company, Philadelphia, 1949), sixteenth edition.
3. Gurd, F. R. N., "The specificity of metal-protein interactions," in *Ion Transport Across Membranes*, H. T. Clarke, editor (Academic Press, Inc., New York, 1954), pp. 246–272.
4. Henderson, L. J., *Fitness of the Environment* (The Macmillan Company, New York, 1927).
5. Klotz, I. M., "Protein interactions," in *The Proteins: Chemistry, Biological Activity and Methods*, H. Neurath and K. Bailey, editors (Academic Press, Inc., New York, 1953), Vol. I, part B, pp. 727–806.
6. Landsteiner, K., *The Specificity of Serological Reactions* (Harvard University Press, Cambridge, 1945), revised edition.
7. Pauling, L., and Itano, H., editors, *Molecular Structure and Biological Specificity* (American Institute of Biological Sciences, Washington, D. C., Publication No. 2, 1957).
8. Waugh, D. F., "Protein-protein interactions," Advances in Protein Chem. **9**, 326–437 (1954).
9. White, A., Handler, P., Smith, E. L., and Stetten, DeW., Jr., "Biochemistry of the endocrine glands," in *Principles of Biochemistry*, Part 6 (McGraw-Hill Book Company, Inc., New York, 1954), pp. 867–997.
10. Wintrobe, M. M., *Clinical Hematology* (Lea and Febiger, Philadelphia, 1956), fourth edition.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

REVIEWS OF MODERN PHYSICS VOLUME 31, NUMBER 1 JANUARY, 1959

1
Molecular Biology and the Physical Basis of Life Processes

FRANCIS O. SCHMITT

Department of Biology, Massachusetts Institute of Technology, Cambridge 39, Massachusetts

PRESENT-DAY research in the life sciences is characterized by unprecedented advances in the biochemical and biophysical analysis of the mechanisms underlying life processes. Biochemistry has passed from the early phase, in which the determination of the chemical composition of complex biomolecules was its major task, through the period of investigation of the detailed processes of intermediary metabolism, whereby "energy-rich" compounds become available to provide the energy for the building, maintenance, and repair of the organismic machinery. Biochemistry is turning its attention more and more to a study of the ways in which this available energy is coupled with physiological processes and particularly with biosynthetic reactions, as in the synthesis of the more critical biomolecules such as the proteins and nucleic acids and their derivatives. In these studies, biochemists require, in addition to the molecules that react in a given process, also an enzyme or enzymes whose role it is to catalyze the reaction in a very specific manner. Biochemistry thus has become to some extent the organic chemistry of enzyme-catalyzed reactions. But it has become obvious also that, although each energy-yielding or energy-requiring reaction in the cell is catalyzed by a specific enzyme, reactions in the cell involve enzymes organized in specific *patterns* or structural arrays. This organization or ordering of enzymes permits cycles of reactions to occur that would be highly improbable or impossible if the individual steps in the reaction were catalyzed by individual enzymes randomly distributed in the cell. The biochemist is now striving to isolate not only individual enzymes required for particular reactions but also specifically organized groups or "assemblies" of enzymes. These are found readily-available in partial systems of cell particulates such as the mitochondria and microsomes that can be isolated from fragmented cells and purified *in vitro* by differential centrifugation. With such subcellular, organized *systems* of enzymes it has become possible to achieve spectacular syntheses in the test tube of proteins, nucleic acids, and other critical biomolecules. With the help of isotopic tracers, it is possible also to determine the ways in which individual groups are shuttled about in these biosyntheses.

Biophysics, which is rapidly developing to the status of a major branch of the life sciences, is following the pattern set by biochemistry, and biophysicists are learning to operate in close harmony with biochemists. Just as biochemists had to determine first the composition and structural chemistry of the complex biomolecules, so in the early development of the field, biophysicists have to determine first, with the aid of crystallographic and physicochemical methods, the detailed configuration of the molecular chains of which the complex macromolecules are constructed. It will then be necessary to investigate the forces between the macromolecules and the smaller molecules in the protoplasmic environment and between various types of macromolecules. By now, it is fully apparent that the interaction properties of the macromolecules crucial for life processes depend upon the highly specific manner in which the covalent chains are structured and integrated with each other by secondary bonds within individual macromolecules. It is equally apparent that these interaction properties also spontaneously tend to form supermacromolecular aggregates whose stability and interaction properties depend sensitively upon their chemical environment of small molecules and ions. These supermacromolecular aggregates are, in fact, not only the active machinery by which muscles contract, glands secrete, and energy is mobilized by spatially organized enzymes in mitochondria, but also the means by which the genetic material is segregated and recombined with mathematical precision and by which the genetic coding is preserved and transmitted to posterity. Such specificity of interaction properties underlies the process by which the coded biochemical and biophysical instructions are passed from the germ cells to the various tissue cells, causing them to develop into the mature organism and to maintain and repair the organism until processes of aging and degenerative disease disturb the nice balance of interactions sufficiently to cause death.

Thus biophysics, like biochemistry, has to reckon with hierarchies of organization and with the properties that are characteristic of systems no less complex than those provided by living organisms at each particular level of organizational complexity, *viz.*, molecular, macromolecular, subcellular, cellular, supercellular, organismic, and superorganismic.

It may be apparent that some of the foregoing is based on prognostication of advances presently "in the works" or still to come, as well as being a statement likely to be made by the historian of mid-twentieth century life science. However, the characterization makes it clear that there are two rather different ways of investigating life processes, both of them highly important. One way is primarily analytical. It is illustrated by the biochemist's search for pure enzymes

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

which specifically catalyze individual reactions. This leads not only to an understanding of the individual reactions, but also eventually to the discovery that these reactions constitute complex cycles of metabolic or biosynthetic reactions. The analytical method is characterized also by the biophysicist's striving to isolate fibrous macromolecules in monodisperse condition and to determine the physicochemical properties and intramolecular-chain configuration of these macromolecules. It is inevitable that such investigations must precede successful attempts to determine how the macromolecules interact to cause biological function. The analytical method has made possible most of the striking advances in the biophysical and biochemical sciences which, together with the parent science of general physiology, have come to be known as "molecular biology."

The hallmark of the alternative approach, which may be called the "organismic" or "systems" approach, is to allow the cell, tissue, organism, or groups of organisms to remain intact in their normal environment and to observe the interaction properties of these intact entities, their adaptation to their environment and their tendency to form still more highly organized entities or systems with new properties, many of which are seemingly unpredictable from a knowledge of less complex systems.

Most of the current advances in biophysical science, and certainly the major portion of the present Study Program in Biophysical Science, involve primarily the analytical approach of molecular biology. This may consist in the isolation and examination of particular cellular constituents at the molecular level; in the demonstration of the role of free radicals or "excitons" in particular cell processes; or in the application of polyelectrolyte theory in an understanding of muscle contraction, to cite a few examples. Actually, the tendency of physicists who have become biologists or biophysicists has been to search for simplified models of complex biological processes and, by a sophisticated study of such models, to discover fundamental new principles in biology. In this approach, the numerous complexities and vast areas of ignorance that lie between the simplified model system and the end biological process under study are characteristically and purposely avoided or neglected. It is recognized that there is a formidable "black box" between the molecular effectors, as studied in the model system, and the final behavior of the cells or organism under study. It is quite possible that this approach may lead to important breakthroughs in biological theory. One of the best illustrations of its effectiveness is the phage-microorganism studies and the modern biophysical approach to genetics generally. This subject forms an important part of this Study Program. These advances have not been limited to genetics, narrowly conceived as concerned alone with mechanisms of inheritance, but also have illuminated some of the basic biosynthetic processes by which the genetic determiners guide cellular differentiation and regulation. Thus, the contents of the "black box" are also being scrutinized wherever possible.

Actually, much of the substance of present-day analytical biology would, in these terms, be considered "black box" science—i.e., concerned with the intermediary detailed mechanisms as well as with eventual "basic" mechanisms. However, particularly in the framework of a Study Program such as this, it is important that the dynamic, adaptive, regulatory, homeostatic properties of organisms and cells be kept clearly in mind, as well as the more readily studied properties of simpler partial systems. For such purposes, textbook descriptions of the "typical" cell are of little help. Rather, the living cells themselves must be studied, and it is the purpose of the first papers in the Study Program to demonstrate some of these properties.

In carrying out my own assignment of discussing the role of cell constituents in life processes, I should like to sketch very briefly the present situation in cellular biology in the light of some of the great accomplishments and controversies of the past. It may be hoped that this not only will provide much needed perspective concerning the "systems" or organismic properties of living organisms but also may stimulate some to consider the vast opportunities in theoretical biology, a field that is almost nonexistent at the present time, at least if judged by the standards of theoretical physics. Of necessity, such a theoretical biology must deal not only with the properties of cellular constituents but also with the properties of the organism as a whole. It must seek to identify and characterize those properties of living systems which, like the quantum of action, are ingrained in nature itself and must be assumed as a principle of operation of the living system.

HISTORICAL PERSPECTIVES

In characterizing the search for the physicochemical basis of life processes, it is pertinent to mention a few concepts and controversies as well as several of the major theories and principles. In these one may trace, from the very beginnings of Greek thought, the thread of the conflict between the atomistic or analytical and the holistic or organismic (molecular biology contrasted with organismic or evolutionary biology).

Epigenesis *vs* Preformation

The principle of epigenesis—namely, that the *potentialities* rather than the miniature structures of the adult organism themselves reside in the egg or sperm and that, in development, the tissues and adult structures arise by an orderly series of structural and chemical changes—was firmly grasped by Aristotle and, in the seventeenth century, by William Harvey. However, in the seventeenth and eighteenth centuries, there arose

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

speculative controversies about the possibility of "preformation"—i.e., the possibility that the egg or sperm contains within it an embryo fully formed in miniature, that in development the miniature organs unfold into the mature form. Many of the great scientists of the day adhered to this patently impossible concept, differing among themselves chiefly as to whether the sperm (Leeuwenhoek, Boerhaave, Leibnitz) or the egg (Swammerdam, Malpighi, Haller, Bonnet) contained the miniature structures. This kind of speculation was terminated by the demonstration by the early embryologists of the progressive manner in which new structures are formed from previously existing simpler systems. The investigations of the continuity of genetic determiners, as well as the pluripotential properties of the cells of the early embryo, were enriched greatly by the ingenious experimental work in developmental biology in the last few decades.

Obviously, a certain amount of basic information must be carried in the egg and sperm, and this is currently regarded as being the "coding" or instructions of how to construct the structures rather than being microeditions or templates of the structures themselves. It must be remembered that, in addition to the genetic coding which must be preserved intact in the germ cells for transmission to the progeny, there is also the coding that goes to every cell of the body and that directs the chemical and structural differentiation of the various tissues and organs during development. For example, certain cells of the developing embryo must be instructed to become differentiated from their fellows as limb buds and these in turn must cause the orderly development not merely of the skeleton and muscles of the limb but also the appropriate innervation from the developing nervous system, as well as the various secondary patterns of organization of the limb such as the color and patterns of feathers, hair, and so on.

Great emphasis in recent years has been placed on the mechanism by which deoxyribonucleic-acid (DNA) macromolecules uncoil, replicate, and recombine, in explanation of the nature and action of the genetic determiners, and on the manner in which ribonucleic acid (RNA) is coded by the DNA and is, in turn, made the code for the biosynthesis of proteins and other complex biomolecules in the cytoplasm. There is, however, little definitive evidence presently available concerning the manner in which this coding is regulated, modulated, activated, and inhibited so as to make possible the adaptive, pattern-directed reactions that characterize the normal processes of development, maintenance, and repair. There is a tendency among biophysicists to take all of this more or less for granted as being primarily a problem of complexity—properties of the black box. But this may well turn out to be as much a part of the essential problem of life as the definitive coding and biosynthetic processes as such. We may, in fact, be dealing here with a bit of the essence of the biological problem generally—i.e., the difficulty in arranging not only to provide primary, secondary, and derivative codes for particular reactions, but also to see to it that the proper substrate and enzyme molecules are ready at hand at the right places and all at the right moment so as to make possible orderly development, regulation, maintenance, and repair. As Delbrück* has indicated, there is still a great gap between the DNA map and the genetic map; probably also between many other complex biological processes and the macromolecular coding that apparently directs the process. The nature of this regulatory, adaptive, feedback mechanism also is discussed in this Study Program, though regrettably not nearly so thoroughly as the importance of the subject demands.

Cell Theory

Formulation of the cell theory in 1838–1840 by Schleiden and Schwann proved to be one of the greatest generalizations in biology, ranking in importance with the theory of evolution, because it made possible the explanation of biological process in terms of the cells of which the organism is composed. The rapid rise of physiology and of pathology as well as an understanding of the special significance of the germ cells followed as soon as the concept of the cell as the unit of life was fully apprehended.

In the century that followed, it became obvious that the idea of the cell as the unit of life must be modified. Any particular cell is what it is by virtue of its relationship to all the other cells in the body. The organism is an ideal federation of cells, each with a high degree of specialization, division of labor, and a corresponding mutual interdependence. It is the organism that is alive in the strict sense of the word, though individual constituent cells may readily be removed and cultivated *in vitro* under appropriate conditions.

Theory of Organic Evolution

Mention is made of this most important hypothesis primarily to emphasize the fact that it was the result of naturalistic observations on *whole* organisms in *whole* communities. It was developed quite independent of the then current tendency to focus attention on the cells and their parts. Darwin was, in fact, singularly misleading and ineffective as an analytical "molecular biologist" (witness his theory of pangenesis and his hypothetical "gemmules"). However, the evolutionary theory strongly reinforced the need to explain the mechanism of inheritance and provided a challenge to molecular biologists for more than a century to come.

Mendel's work provided a beautifully clear set of basic relationships. But, unfortunately, they were soon

* M. Delbrück, "Atomic physics in 1910 and molecular biology in 1957," a lecture delivered at the Massachusetts Institute of Technology in connection with the Karl Taylor Compton Lectures by Professor Niels Bohr.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

forgotten and had to be reintroduced at the turn of the century.

Theories of the Fundamental Structure of Protoplasm as the Physical Basis of Life

Soon after the formulation of the cell theory, attention was focused upon the protoplasm (so called by Purkinje in 1840) as the physical basis of life.† The search was on to discover what is actually alive in the cell. With the rapid development of optical instrumentation, methods of "preserving" (fixing), sectioning, and staining tissues, there was ushered in the golden age of descriptive morphology. Investigators were in hot pursuit of the "living subcellular particles," the ultimate living units.

To discoveries made during this very active period one owes the development of the knowledge of some of the most fundamental cell phenomena. In this active expansion of descriptive cytology from 1870 to 1890, three general theories were proposed for the fundamental structure of protoplasm as the physical basis of life, as follows:

1. *The Fibrillar Theory ("One-Dimensional" Arrays)*

According to this view, proposed most succinctly by Flemming (1882), "The essential energies on which life depends have their seat in fibrillae." Heidenhain's book *Plasma und Zelle* abounds in examples of fibrous structures in cells (muscle, mitotic apparatus, cilia, flagella, sperm tails). Two fibrous types were distinguished: the unbranched (filar) and the branched (reticular). The fibrils are immersed in a clear "hyaloplasm" (Leydig).

2. *Alveolar or Foam Theory (Two-Dimensional, Membrane-Limited Structures)*

The chief proponent of this theory, Bütschli (1878–1892), believed that protoplasm is a diphasic system containing a continuous clear, viscous fluid (hyaloplasm) and a discontinuous phase of microsomes (*sic*) which are probably lamellar in life. In this theory, fibers were ignored. However, since fibers undoubtedly do exist, Strasburger (1892) proposed a compromise. Protoplasm contains two plasms: a "trophoplasm" which is membranous or alveolar and whose function is metabolic and nutritive, and a "kinoplasm" which is fibrillar and which mediates movement or contraction of any kind.

3. *Granule Theory*

Between 1880 and 1890, Altmann exhaustively described certain granules ("plastidules") in protoplasm which he regarded as the organic units of the cell, as the cells are the organic units of the organism. He considered the granules to be "living elementary organisms (bioplasts)." In 1890, he coined the phrase "*omne granulum e granulo*," paraphrasing Virchow's earlier (1855) dictum "*omnis cellula e cellula*."

However, as is now known, the cytologists were being led astray by their insistence that the structures they observed in fixed tissues were present *as such* in the living cells. A halt was called, particularly by W. B. Hardy, at the turn of the century, to the speculative race by descriptive morphologists to discover the physical basis of life, through observation of structures in fixed and stained tissues. He showed that many of the structures claimed by cytologists to be the physical basis of life are, in fact, coagulation artifacts. He was able to imitate many of them in artificial systems subjected to precipitation under controlled conditions.

J. Loeb's trail-blazing work on the physicochemical and colligative properties of proteins helped remove the needless complexity with which the early colloid enthusiasts had cloaked their studies of protoplasm.

The startling work of Langmuir, Harkins, Rideal, Adam, and Gorter on molecular orientation in mono- and multi-molecular films laid a basis for an understanding of the properties of interfacial films and factors stabilizing polyphasic systems. Studies of the paracrystalline, mesomorphic state, beginning with the work of Lehmann on fluid crystals, led to an understanding of molecular ordering in anisotropic systems.

The return to the study of *living* cells by means of micromanipulation (Peterfi, Chambers) and by tissue culture (Harrison, Levi, Weiss) helped offset the eclipse into which descriptive morphology had been sent by the skeptical, physicochemically minded physiologists, and revealed new and important properties of cells and tissues.

With the introduction of electron microscopy, and particularly since the development of techniques for examining tissues in ultrathin (100 to 500 A) sections by high-resolution (*ca* 10 A) electron microscopy, we seem to have entered upon a new golden age of *ultracytology*. From a study of tissues fixed by only a few types of fixatives (because of limitations imposed by plastic embedding and other preparative techniques as well as because of the intrinsic qualities of preservation of the cells), descriptive papers in ever-increasing number (and length) are being produced as electron-optical equipment is made easy for the nonphysicist to operate and is being installed in more and more laboratories the world over.

As in the analogous period of development of light microscopy in the last century, the new flood of electron-optical observations has brought with it important new discoveries concerning the structure of protoplasm and its subcellular, "submicroscopic" particulates or organelles. There is also danger that one may again be misled in matters of interpretation because of the fixation

† For an excellent review of these early studies, the reader is referred to the book of E. B. Wilson, *The Cell in Development and Heredity* (The Macmillan Company, New York, 1925).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

artifacts that are now being observed at a hundred times the magnification available to the morphologists of the last century. However, the danger of this is greatly reduced because other biophysical and biochemical advances make it possible to obtain independent evidence concerning the molecular organization of protoplasm.

CURRENT CONCEPTS OF CELLULAR ULTRASTRUCTURE IN RELATION TO FUNCTION

Electron microscopy has revealed highly significant new facts about each of the three types of protoplasmic structures that figured prominently in early theories of histology: the granular, the membranous, and the fibrous structures. For present purposes, it suffices to indicate briefly the upshot of the more significant of these discoveries.

By appropriate fragmentation or maceration of cells, followed by differential centrifugation and purification, it is possible to prepare, in separate fractions, certain of the cellular organelles believed vital by the early cytologists. Their biochemical properties and ultrastructure also may be studied in one and the same preparation.

In this way, it was shown that mitochondria are the "power plants" or energy source of the cell by which, through oxidative phosphorylation, phosphate-bond energy is made available through the splitting of adenosine triphosphate (ATP), the metabolic fuel of the cell.

The most characteristic aspect of mitochondrial organization is the internal layered structure. Biochemical studies show that both the outer enclosing membrane and the internal lamellae contain the enzymes required for the citric-acid cycle. From the work of Lehninger and others, it seems possible that the various individual enzymes involved in such biochemical cycles occur as clusters or "assemblies" anchored in or upon the membranous structures. The direct demonstration of such assemblies and of the molecular organization of the layers now becomes a challenging problem in electron-optical technique. Thus, although mitochondria play a vital metabolic role in protoplasmic function, they are far from being the elementary units of life that Altmann pictured them to be.

In the presence of appropriate substrate molecules, cofactors, and activators, fragments of macerated cells have proved capable of rather spectacular biosynthesis of peptides, steroids, and other complex biomolecules. Apparently, sufficient amounts of enzymatically active cell surface are retained in such preparations to produce biosynthesis. This has proved to be a very valuable biochemical tool particularly when used in conjunction with isotopic tracers.

Similar results were obtained with the so-called "microsome" preparations from fragmented cells subjected to differential centrifugation. The microsomes occur in the fraction lighter than that containing the mitochondria. Electron-microscopic examination shows these microsomes to be fragments of the membrane-limited structures in the cell ("endoplasmic reticulum," "cytomembranes," "ergastoplasm") which, when associated with RNA-rich particles ("ribosomes"), are thought to constitute the biosynthetic center of the cell. Electron-microscopic examination of very thin sections of actively synthesizing cells suggests that the cytoplasm may be divided into two phases by such membrane-limited structures. One phase is continuous with the space between the nuclear membranes and perhaps also with the extracellular fluid by means of micropores in the limiting cell membrane. This phase contains a relatively homogeneous fluid material which doubtless contains biosynthetized materials. The second phase contains the mitochondria and the RNA-rich granules (ribosomes), which are the biochemical and structural apparatus of biosynthesis. Thus, there may actually be two plasms in cytoplasm (one of which might perhaps be called trophoplasm) as adherents of the alveolar theory supposed. However, the membranes that separate the two are themselves metabolically very active ones which, by partitioning off regions of the cytoplasm, facilitate the biosynthetic process.

Other membrane-limited structures are also of fundamental significance. Of these, perhaps only two need be mentioned: the Golgi system and the limiting envelope or plasma membrane of the cell.

The Golgi "apparatus," which was considered by cytologists to play an essential role in secretion processes, came under severe criticism as a possible fixation artifact. However, electron-microscopic observations indicate that it is a multilayered structure frequently seen in the secretogenous zone of the cell. It is supposed by some to be the "wrapping and packaging" department of the cell in which secretion granules and other particulates may be enclosed in a membranous wrapping which, by fusion with the cell membrane, allows the granules to traverse the plasma membrane.

The plasma membrane serves not only to enclose the cell and to direct the molecular traffic into and out of the cell but also, presumably, to mount the biochemical mechanism by which solutes, such as sodium ions, may be transported or "pumped" against an activity gradient. Electron-microscopic observations indicate that such membranes have a structure similar to that of the typical unit membrane in cytoplasm generally. This consists of two dense lines separated by a less dense area. The total thickness is 60 to 70 A and may consist of a bimolecular layer of mixed lipids with a monolayer of protein or other material on either interface. As in the case of mitochondrial membranes, it remains to be seen whether or not improved techniques can reveal the presence of specialized, possibly enzymatic, structure having a role in transport or other biochemical processes mediated at the cell surface.

One comes finally to the fibrous types of arrays which are ubiquitous in protoplasm, being particularly con-

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

spicuous in the connective tissue, in the mitotic apparatus of dividing cells, and in tissues such as muscle where rapid and reversible interaction of the macromolecular lattice causes contraction or produces tension.

Since the early pioneering work in the twenties and thirties, particularly that of Astbury, many fibrous proteins have been isolated and investigated with respect to their structural and physicochemical properties. It has been found that these elongate macromolecules manifest a high degree of specificity in their ability to aggregate in ordered arrays, to change the pattern of interaction with change in the chemical environment, and possibly to change the intramolecular configuration (helical, supercoiled, and uncoiled, random chains). This matter is dealt with in detail in this Study Program in the case of certain proteins such as collagen and certain muscle proteins which manifest a specific band structure in the electron microscope and by means of which their aggregation states can be determined.

The production of fibrous lattices in cells and tissues apparently depends upon such builtin properties of the macromolecules. There are no microworkmen to go about in the cells joining up the proper molecular beams of the scaffolding according to some predetermined blueprint! Rather, after the elongate macromolecules have been synthesized, and possibly activated by preliminary enzyme action, they automatically polymerize and aggregate with each other according to their builtin specificity, to form the structure or microlattice appropriate in each protoplasmic situation. Thus, the genetic code may be thought of as arranging for the biosynthesis of the complex macromolecules of many kinds, but interaction of the macromolecules under specified conditions occurs spontaneously. In many instances, for example, characteristic fibrogenesis can occur *in vitro*.

Thus, fibrous arrays are essential, not only to provide structural and contractile elements but also, as in the case of the nucleic acids, to provide a means of arranging specific linear sequences of determiners corresponding to particular chemical codes. However, even the most enthusiastic exponent of DNA and viruses as linear codes, would hardly ascribe to them the properties of life such as was suggested by Flemming and the adherents of the fiber school.

Thus, as the specialized biophysical and biochemical functions of morphologically identifiable components, such as the mitochondria, ribosomes, Golgi system, and other membrane-limited structures, come to be better understood (chiefly by a study of partial systems *in vitro*), the importance of the organization of the cell as a whole becomes more and more evident. It is this highly coordinated temporal and spatial ordering that provides the dynamic, regulatory, homeostatic, and adaptive properties of the cell which are the essence of life. Somehow, the biosynthetic capabilities of the cell are regulated and geared into the energy-yielding reactions of metabolism so as to produce the substances that may be required not just locally within the synthesizing cell but by cells at remote loci in the organism. One has only to think of the complex ordering that controls the processes of growth and development in which the individual cells seem to be governed by master plans or codes (morphogenetic "fields" of general or local specificity) but in which the individual cells obviously must be constantly producing substances which make it possible for them to react meaningfully to the general and local fields. Indeed, the very existence of the field depends upon such processes in the individual cells. Yet the master plan (genetic coding) is characteristic of the entire organism as a whole, not only during development and growth but in all of the processes of aging.

One sees illustrated here in especially impressive fashion a general property that seems to characterize living systems at all levels of complexity and that has in it an indeterminacy which is of fundamental significance. Somehow, the appropriate reactions occur at the right places and at the right time, in individual cells and in the organism as a whole, to produce the eventual result preordained in the coding of the DNA transmitted in the germ cells.

It is easy to see, therefore, why it is difficult at the present stage of scientific development to deal with problems of this nature at the molecular level with precise methods of physics and chemistry, and at the same time to characterize with equal precision the ordering and the reactions that govern the over-all "directive" behavior of the organism as a whole.

To a certain degree, the difference in mode of attack is revealed in the various papers in this Study Program and is reflected also by the professional orientation of those whose efforts are primarily in the area of "molecular biology" as compared with those dealing with the larger, systems aspects of cell, tissue, and organismic function.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

2
Cellular Dynamics

PAUL WEISS
The Rockefeller Institute, New York 21, New York

BIOLOGY is now in an exciting phase marked by the confluence of different disciplines of research in the attack on focal problems. But the enthusiasm raised by the spectacular results of combined physical and chemical approaches to biology sometimes has outraced people's ability to keep pace conceptually with the technical developments. As a result, we frequently try to fit our questions to the very limited answers which our fragmentary knowledge has been able to provide, instead of boldly facing the much broader questions posed by living systems and phrasing them in such a way that still more penetrating answers may be obtained in the future. In order to do this, one needs to focus on the real living objects, rather than on the somewhat fictitious and oversimplified models that one is prone to formulate, as intended targets for physical and chemical attack. Models are necessary, but they must bear more than a coincidental resemblance to the real object if they are to serve as meaningful aids to analysis.

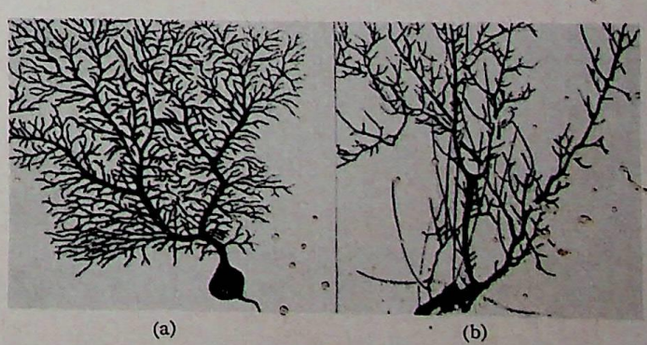
As the result of this extensive use of overly simple models, notions about the cell have become at times slightly vague and unrealistic. It would be presumptuous in one single chapter to try to do more than to just give a few illustrative examples of what the real cell is like. The best that can be hoped is to show the change that has occurred in our thinking about the cell from the static to the dynamic—that is, from static organization to organized behavior. Much of the knowledge of what the cell is has come from ruling out erroneous conceptions of what it is not. Progress has come from narrowing the margin of error. By being exposed to a few examples of the living cell in action, the reader can judge for himself whether or not his mental picture of the cell corresponds to the real thing.

The first example deals with one of the most prominent characteristics of the cell—its shape. Figure 1(a) shows a textbook picture of a particular cell found in the cerebellum. One sees that the cell body has elaborate ramifications. This is the way one usually learns about a cell—through pictures in a book; and since the picture looks the same in all of the thousands of copies of a textbook of microscopic anatomy, one forms the notion that all such cells are like tin soldiers stamped out according to a standard pattern. Thus, the mental habit of cell form as something static and rigid becomes engrained. The truth is, however, that no two cells are ever strictly alike, nor is any one cell quite the same at different times of its life history. It is this history which a static textbook picture fails to reveal. To stress this fact, Fig. 1(b) shows by comparison a Chinese brush drawing of some shrub. Here common experience tells us that the bush has not been stamped out in the shape in which one finds it. It has grown into that shape from seed. So, what one sees as pattern in the shrub is merely the residual record of prior activities of that particular protoplasmic system. In other words, shape is simply an index of antecedent processes by which that shape has come about.

Something else is lacking, however, in both of the pictures besides the account of prior events. The objects are portrayed against a blank background as if they were in a vacuum. Again, in the case of the plant, one knows from daily life how vital the invisible air is for its existence as a provider of chemical necessities.

Now, in the case of the cell, the medium is involved even more intricately: it provides not only chemical components for nutriment, but also a physical framework that integrates the separate cells into a structural continuum. The existence of this continuum usually remains unrecognized because staining techniques deliberately leave the substratum out of sight. Yet, to the classical morphologist only seeing was believing and, what is worse, not seeing amounted to not believing. This attitude is undergoing radical change.

Processes as such are not visible. What is visible is a constellation of elements at different stages of the process. Visible form, a pattern at any one stage, must be viewed as the product of antecedent formative processes. The cell thus appears as a system of highly complex, but ordered, molecular populations grouped in a hierarchy of supramolecular complexes, in constant interactions among themselves and with their environment (that is, the space beyond the cell border), leading to features, some permanent, others transitory, the visible expression of which is recorded as shape. If the environmental conditions are reasonably constant for a group of cells of the same type, the behavioral history of the latter will be reasonably similar so as to end up



(a) (b)

FIG. 1. Comparison between a nerve cell (a) and a plant (b).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

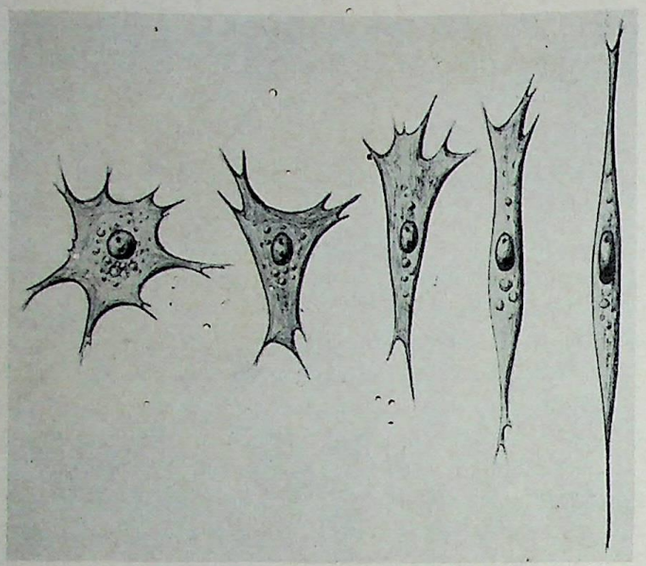


FIG. 2. Samples of shapes assumed by cells of the same connective-tissue cell strain in tissue culture.

with reasonably similar and classifiable shapes of the sort that have made it possible for sciences of microscopic anatomy and microscopic pathology to develop. As soon as there is a change in the conditions, the behavioral response of the cell likewise changes and the familiar shape derived from normal standard conditions ceases to be a diagnostic sign.

A most dramatic illustration of this situation is seen when cells are taken out of their normal site in an organism and transferred into an extraneous medium in tissue culture. As an example, a time-lapse phase-contrast microcinematograph* of human-liver cells spread on glass in horse serum (film made in my laboratory by A. Cecil Taylor and Albert Bock) shows up impressively the lack of fixity and the incessant reshuffling of cell content and contour. No static description can do justice to this vivid record of ever-changing activity. These liver cells in culture look quite different from those one would be used to seeing in stained sections through an intact liver. Except for shape, however, they still possess most of the essential properties of liver cells. In a third type of setting, for instance, in suspension, they would assume still other shapes.

Thus, one realizes that there is no way of getting a fully valid description of a cell except by studying its behavior under as wide a spectrum of conditions as is feasible. Cells of different kinds behave differently. While the transfer to tissue culture alters their morphological expressions markedly, they do retain their constitutional distinctions of behavior.

In conclusion, one is led to the thesis that cell shape is the result of a distinctive behavioral reaction of a living cell to its environment.

To make this concrete, consider a specific example. A colony of fibroblast cells cultured in dilute blood plasma yields a wide spectrum of shapes. The same cell can appear in any of the series of forms pictured in Fig. 2, ranging from the bipolar spindle at one extreme [Fig. 2(e)] to the multipolar star at the other [Fig. 2(a)]. The shape thus depends upon the number of directions in which the cell border shows radial extensions. The tips of these extensions are the active mobile organs of the cell. They push outward and thus distort the originally rounded surface of the cell. Evidently, if there are only two processes tugging in opposite directions, the cell body is drawn out between them into the shape of a spindle [Fig. 2(e)]. If there are three major protrusions, the cell assumes a tricornered shape [Fig. 2(b)], and with even more processes along its circumference it approaches more and more a star shape [Fig. 2(a)].

Must one accept this spectrum merely as a given descriptive fact, or can it be explained causally in the way physical systems are treated? The answer is that, to a certain extent, the whole series can be expressed in terms of a single function derived from a study of cellular behavior. Cells do not live in a structural vacuum as is the illusion created by standard histological preparations, which, by stressing only those features which happen to be stained, obliterate the structural continuum within which the cells reside. In tissue culture in a blood-plasma clot, for instance, this continuum is provided by a network of fibrin fibers—aggregates of molecular chains of varying diameter from submicroscopic to microscopic dimensions, the meshes of the network being filled with serum (Fig. 3). It is in this fibrous jungle that these cells live and move, applying themselves to the interfaces between the fibers and the liquid medium as to a trellis. As was mentioned before, the shape of these cells is determined by the number of protrusions from their surface. One can go one step further and prove that the number of such processes, in turn, is a function of the fibrous constitution of the medium.

FIG. 3. Electron micrograms of plasma clots coagulated at different *p*H values (from left to right: alkaline, neutral acid).

* The motion pictures referred to in the text were shown at the Study Program in Biophysical Science in conjunction with the lecture on which this article is based.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

The relevant interaction is between the cell surface and the fibers in its microenvironment. To understand such surface reactions, one must give up in the first place the outdated notion that the cell surface is a sort of static cellophane-like bag. This may be true of some specialized cell types, for instance, the cellulose membrane of a plant cell, the capsule of a bacterial cell, or the envelope of a red blood corpuscle. But in most types of cells, the surface is far from stable and is by no means of identical composition and state all over the cell. In tissue culture, this state of disequilibrium manifests itself in the continual thrusting forth and withdrawal of surface processes at the expense of cellular energy, showing great variations of the contractile force along the surface. Temporarily weaker points along the surface thus become outlets for thrusts. Such microleaks or "herniations" may occur at random or they may be determined systematically by outside factors, of which one of the most important is the encounter of a fibrous-liquid interface with the cell surface. The fiber contact, in a sense, pricks the cell surface locally. The strength of the resulting strain can be shown to vary with the size of the fibers. Hydrodynamic, viscous, and elastic competition for outflow favors fibers which have a larger diameter (Fig. 4). Consequently, the prevalence of a few major protrusions over minor ones may be expected to be the greater, the larger the average fiber size is in the medium.

These predictions have been tested (jointly with B. Garber) by culturing cells in plasma clots containing fibrin fibers of different average dimensions. The average diameter of such fibers is a function both of *p*H (Fig. 3) and of plasma concentration, larger fibers being formed at either lower *p*H or higher plasma concentrations. It actually was found, in line with expectations, that the ratio of bipolar cells (few processes) over multipolar stellate forms (many processes) increased as a steady linear function as the plasma concentration was raised

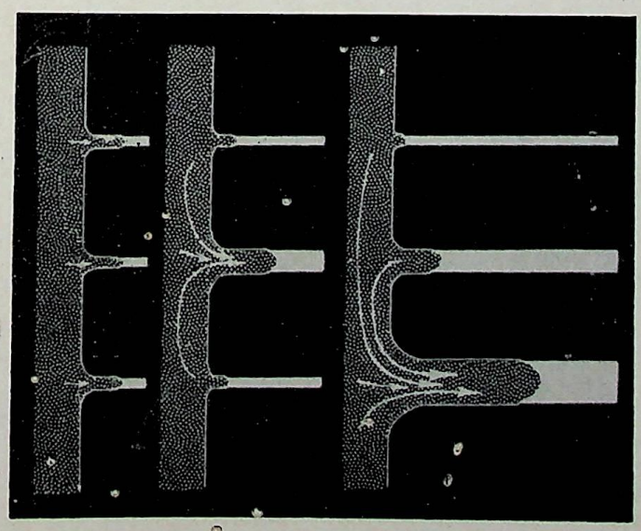


FIG. 4. Microherniations of cell content at intersections of fibrin fibers of various sizes with cell surface. Arrows indicate protoplasmic outflow.

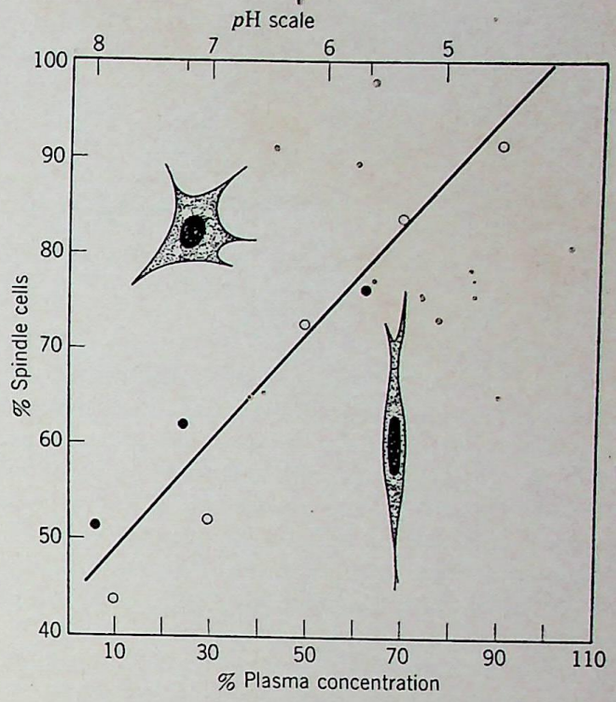


FIG. 5. Distribution of cell shapes in populations cultured in plasma clots of different average fiber sizes.

or as the *p*H during clotting was lowered (Fig. 5). Other criteria of cell shape, such as the ratios of length over width of the cell bodies and of the cell nuclei, showed correspondingly systematic changes. In other words, the whole gamut of shapes displayed by this particular cell strain could be written in a single formula derived from insight into the mechanisms by which deformations into one or another shape come about. The point to stress is that one gets further by studying realistically the formative process rather than by dwelling upon pictorial samples of forms already achieved.

At the same time, it must be stressed that the formula is a probabilistic one, for to predict just what any individual cell will look like is impossible because of the accidental nature of the details of its surroundings. The microclimate and the microenvironment of the individual cell are unique, unknown in each particular instance, and this establishes a certain degree of variance for the actual expressions within each cellular system which is built into its nature.

In a different medium, the same cell strain would give different responses. For instance, these cells, when suspended in a liquid medium without interlaced fibers, would manifest the inequalities along their surfaces by blunt herniations rather than by the pointed protrusions noted along filaments. As a result, such cells appear to be blistering and boiling along their surfaces as can be seen in cinematographs of unattached single cells, and is well known from the loose cells in the late stages of cell division. Conversely, cells of different tissue types would show morphological responses different from the strain just exemplified. The more a cell is given to producing internal structures that serve as a cytoskele-

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

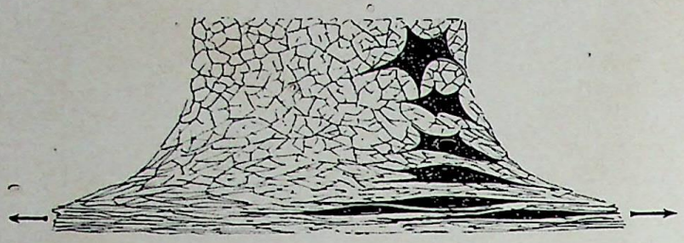


FIG. 6. Effect of regionally varying degrees of tensions on the organization of a fibrin network and, through it, on the morphology and orientation of enclosed cells.

ton—such as in sperm cells, higher forms of protozoans, or muscle fibers—the less will its shape be codetermined by physical constellations in its environment, although even in the most extreme case the polarization of the axiated system of the cell is presumably still a response to external gradients or to other inhomogeneities of the environment. There is an enormous task laid out here for future detailed investigations on the physical factors involved in cytogenesis and morphogenesis.

As a further example of the complexities involved, one can again cite experiments with connective-tissue cells in tissue culture. A long time ago, I became preoccupied with the role of external factors in guiding the oriented movements of cells. In 1927, I succeeded in orienting cells in tissue culture by applying stretch to the blood-plasma clot; the cells assumed a common orientation along the lines of tension. The analysis of this phenomenon over the years has led to the following summary conclusion: The primary effect is on the orientation of molecular chains which become aligned in the direction of the stretch. When cells are contained in such a medium, their shape—that is, behavioral deviation from a sphere—simply reflects the degree of structural organization of the underlying submicroscopic fibrous network. Where all fibers run parallel and, therefore, where one direction only is open to the cells, they naturally become bipolar, all pointing in the same direction (Fig. 6). With decreasing amounts of stretch and correspondingly less rigorous orientation of the fibrin, cell forms grade over from the strictly determined spindles at one extreme to the only probabilistically describable cells of the multipolar sort mentioned in the earlier example. In the present case, however, the average shape does not vary from clot to clot in accordance with the average constitution of the clot, but varies locally within the same clot in accordance with a systematic variation of an extrinsic factor, namely, stretch.

However, the immediately relevant thing for the orientation of the cells is the orientation of the fibrous pathways which they are bound to follow, and it is immaterial that, in those earlier experiments, the agent for producing such fibrous orientation was stretch. Even though the normal organism frequently uses stretch in that capacity, other forces also are at work to produce oriented fiber patterns, which likewise act as guides for cells. The inside of a fish scale, for instance, contains layers of collagen fibers beautifully arrayed in parallel lines, and when loose spherical cells are deposited on such a sheet they immediately assume spindle shapes, with the axes strictly aligned along the fibrous substratum (Fig. 7). Therefore, to carry the analysis further, one must study why a round cell on a linear track becomes correspondingly deformed.

FIG. 7. Elongation of erstwhile-round connective-tissue cells placed on a sheet of parallel collagen fibers from the interior of a fish scale.

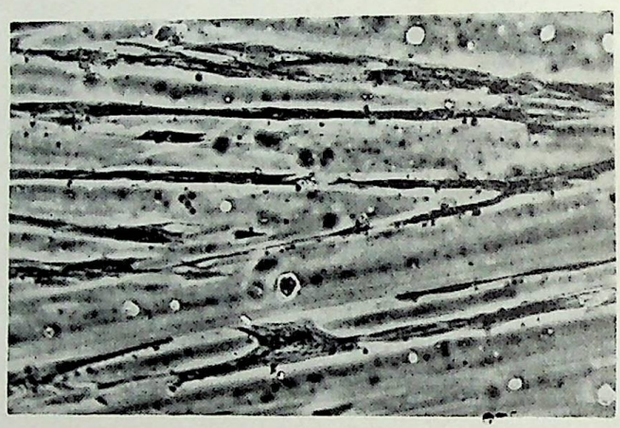


FIG. 8. Phase-contrast photomicrogram of cells which have become elongated along streaks of silicone paste on glass (round white and dark splotches are clusters of cells over nonadhesive silicone where they have been unable to attach themselves).

A clue to the mechanism operating in this case may be obtained from the following experiment. A glass slide first is streaked with silicone paste or with cholesterol. A loose cell suspension then is placed on top in an appropriate liquid medium. As the cell surface does not adhere to cholesterol but does adhere to glass, the cell becomes drawn out along the striplets of bare glass; hence, it is oriented parallel to the streaks (Fig. 8). Shape is determined here by the differential adhesion of different parts of the cell surface to the substratum. An even more impressive example of this kind is provided by cells which have been seeded out on glass scored with a microlathe (experiments performed with A. Cecil Taylor). These cells become deformed from a spherical to an elongate shape in the direction of the microgrooves (Fig. 9).

What then precisely is the mechanism of this response? To understand it, consider an older experiment in which spindle-shaped cells were made to spread out

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

flat when they were forced to splash against a smooth surface from above. Such a surface offers the cell innumerable directions in which to radiate at the same time. Thus, the periphery of the cell actually expands concentrically, flattening the cell into a disk (Fig. 10). This transformation is accompanied by profound changes in function—the cell becomes phagocytic—and in the distribution of intracellular materials, further emphasizing the intimate dependence of chemical activity on the physical constellation in a cell. Similarly, a freshly isolated cell set down on scored glass spreads out at first in all directions, but only those sectors of the cell border which lie in a linear direction of the substratum retain a foothold and continue to advance. The other sectors are retracted and become consolidated as cell flanks.

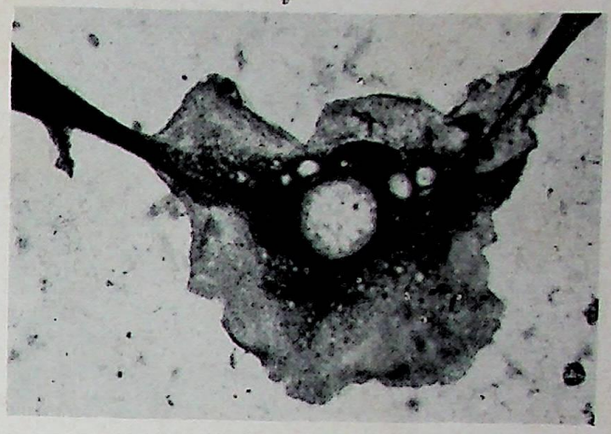


FIG. 10. Spindle cell expanding into a flat disk on impact with glass surface.

Each cell surface thus acquires radically different properties at the ends and along the flank. The two "engines," one at either end, remain active and by extending in opposite directions simply draw the cell out

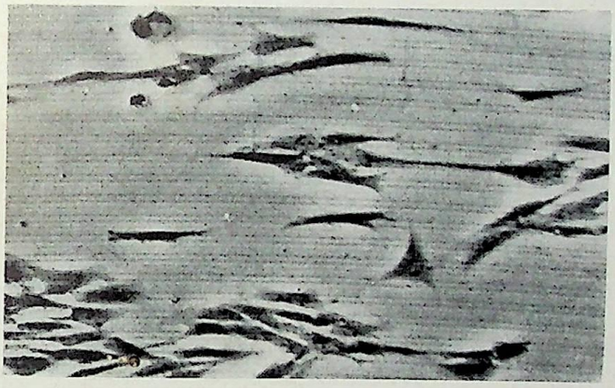


FIG. 9. Cells which have become elongated along the microgrooves of scored glass.

into an elongate form. Such elongation again has further consequences for the cell. For instance, the mitotic spindle preparatory to cell division mostly will become aligned with the long axis of the cell so as to make elongation a major factor in the orientation of cell growth. But because of the opposing tugs, to which this elongate cell is subjected from its two active ends, there is no net forward movement. Such a cell simply shuttles back and forth about a stationary position, comparable to Brownian motion, depending upon which one of the two ends happens to have the upper hand at the moment.

Thus, although the foregoing has brought some deeper understanding of cellular orientation, it tells nothing about the mechanism of cellular locomotion, which remains one of the basic unsolved biophysical problems. Something is known about those cases in which cells have special locomotor organs, such as cilia or flagella, but when it comes to cells moving with their free, unstable surfaces devoid of structural specializations that could be related to motility, ignorance is profound. Cyclosis in plant cells, the gliding of slime molds, the shift of a sheet of skin to cover a raw wound, the invasion of tissues by metastasizing cancer cells, or the penetration of leucocytes through capillary walls to converge on a focus of infection—in none of these cases is it known just how the cell achieves these movements, except that it is suspected that gelation-solvation cycles or contraction-relaxation alternations may somehow be involved. This is one of the most neglected areas of physical approaches to biology. Not only is the mode of locomotion shrouded in ignorance, but also there is equal uncertainty about the reasons why a free cell, which can extend in many directions, often advances steadily in one direction to the exclusion of others. As was just said, a cell left to its own devices in an isotropic environment strays at random with no net dislocation.

There have been many theories and speculations about the directive movement of free cells. Once again, one can illustrate how progress has come from eliminating among such competing concepts the ones ruled out by factual analysis. Turning again to the sample object of cells in tissue culture, contrary to the scattered population of stationary isolated cells discussed before, the cells of a solid fragment of tissue explanted into culture behave quite differently. They move in droves from the explanted piece into the empty medium, giving rise to the well-known phenomenon of "peripheral outgrowth." Why do they move centrifugally? For a long time, I had considered this question synonymous with that of cell orientation and had invoked the fibrous guide rails as explanations of "oriented movement." But, as was just explained, orientation and displacement are two different things, and for the displacement there has been no crucial explanation. "Tropisms" and gradients of various kinds have been proposed to explain the phenomenon. It has been assumed that cells respond either positively or negatively to differential concentrations of hypothetical "attractive" or "repelling" substances emanating from point sources, even though it has never been possible to demonstrate just how a cell could translate such directional cues into actual convection towards or away from the source. Recent observations on our tissue-culture

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

strain have, however, turned up a wholly different story, and it is this.

The only way to get locomotion of a cell with two motor engines at opposite ends is to stop one of the engines, at least for a while. Then the remaining engine can tow the cell away without opposition. This is precisely what happens in tissue cultures whenever free tips of two cells make accidental contact with each other: the colliding ends become temporarily paralyzed. A wave of retraction runs over the affected processes, which become partly detached from the substratum, and the two cells thus come under the exclusive pull from their remaining motile ends facing away from each other. Thus, they move apart in opposite directions. After some time, the paralyzed ends gradually recover their motility and again take hold on the ground, so that the cells are stalled once more, but at a greater distance from each other.

Extrapolating this process to a cell population with a gradient of densities, such as a tissue explant, it is evident that, of a pair of cells moving apart, the one shifting peripherally has a lower probability of encountering another cell than has the one which moves toward the explant. Statistically, this leads to a prevalence of outward migration even though cells are actually free to move in any direction. Eventually, a situation obtains in which the cells have become so widely spaced that random collisions are no longer likely; at this point further migration ceases. Thus, the only gradient which plays any role is a gradient of population density. The phenomenon is formally comparable to the diffusion of molecules from regions of higher to lower concentration, except that one deals with the random collisions not of molecules but of complex entities which may be treated as units for the purposes of description. Again, a close observation of the behavior of the real living object has brought answers far more concrete than what could be anticipated from such generalities as "attractive" or "repulsive" forces between cells. Parenthetically, it should be stressed that the type of contact separation between cells mentioned above is characteristic of the species of connective-tissue cells here described, as well as of several similar strains, but it does not apply to other cell strains, for instance of the epithelial variety, where reaction of two cells on contact can be just the opposite—namely, their drawing much closer together, provided they are both of the same kind. Further details on this behavior are given by Weiss (p. 449).

As still another instance of the dangers inherent in dealing with the living cell in terms of broad generalities, one may once more cite the behavior of cells of diverse shapes observed in plasma clots of different *p*H or concentrations. As was reported before, such cells have a variable number of processes, each of which now may be thought of as a train behind an "engine." In such cultures, it is possible to calculate the average rates of advance of the various cell types in a given direction by dividing the total distance spanned in a given period by the time elapsed. Since such cells move neither steadily nor in a straight course, however, any such average value for "rate of advance" is wholly unrepresentative of the true velocity of cell motility. As plasma concentration is increased, the cells tend to become bipolar. This means, inevitably, that the average rate of locomotion increases simply because the course is less tortuous and the cell is stalled less frequently by simultaneous divergent pulls from multiple processes to which multipolar cells are subjected. Consequently, in the lower concentration range, where most cells are multipolar, there is a progressive increase of the average rate as the number of processes declines toward the liminal value of two (Fig. 11). Once the great majority of cells have attained bipolarity, the average rate of advance remains constant, expressing more nearly the true velocity. By comparing Fig. 11 with Fig. 5, it can be seen that, at a plasma concentration above 50%, where the "rate" curve levels off, more than three-fourths of the cell population are actually bipolar. Just as one cannot tell the true speed of a railroad train if the times of departure and arrival at the terminal are the only data available and the frequency and duration of station stops on the way are not taken into account, so the mere establishment of the average rate of locomotion of a cell under various conditions has little practical meaning. Yet, the literature is full of examples in which such average rates have been used to assess the effects of a variety of agents or drugs on cells, without due attention to the effects of these agents on the medium, which may alter the whole setting in which

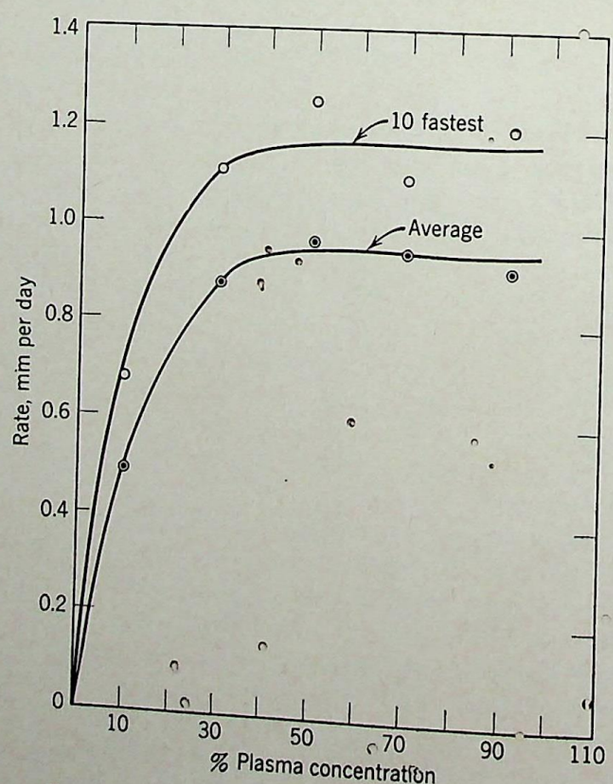


FIG. 11. Rate of progress of connective-tissue cells in plasma clots of different constitution.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

the cells move. Without knowledge of these effects, no meaningful comparisons can be made. This illustrates some of the hazards of operating with *average* quantities when one deals with systems such as cells, the composition and behavior of which are inhomogeneous in space and time.

The foregoing discussion has illustrated how a controlled modification of the medium can indirectly modify cell morphology and behavior, including locomotion. It has explained the orienting effect of tensions on the fiber systems of the medium; how such structures evoke conforming organization in the cell population residing in them; and how cell-to-cell interaction and, in the last analysis, population dynamics govern cell locomotion. In this chain of events, an outside experimenter applying tensions to a culture medium appeared as the primary agent. This, of course, immediately raises the question as to what factors serve this organizing function within the living body. The sole agency of the body is its own cells and it was comforting, therefore, to find that cells, by their own activity, can create the type of orderly structural patterns in the intercellular spaces which had been imitated crudely by extraneous tensions. The cells engender in their own environment physical conditions and orderly restraints which then in turn play back on them as guides and regulators of their own behavior. Thus, a further step of complexity is added to the picture. There are innumerable examples, mostly poorly understood, all showing how, through an enormous variety of mechanisms, the same basic principle is served; which is that the cell population, through its products and interactions, sets up conditions modifying the behavior of the enclosed cells, and thus often leading to new settings and interactions which may cause further alterations of the cells, and so forth, in sequences of interactions of ever increasing complexity.

To illustrate this, consider a piece of tissue embedded in a fibrous network and let the boundary of the tissue

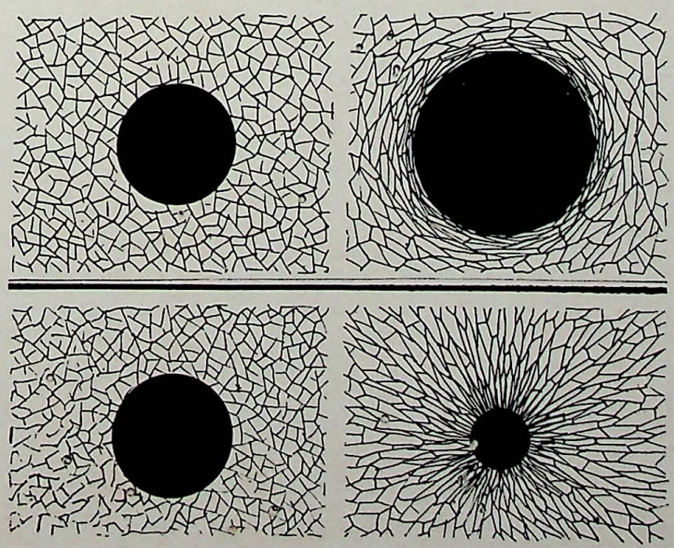


FIG. 12. Effect of an expanding (top) or contracting (bottom) center on the architecture of the surrounding meshwork.

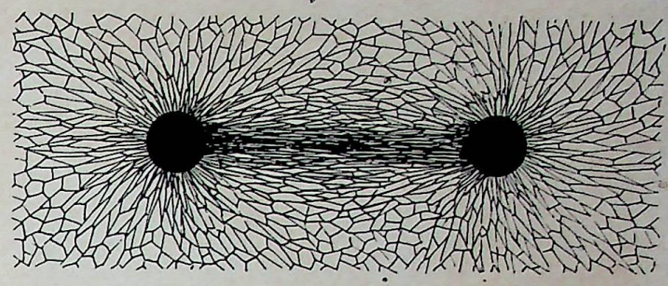


FIG. 13. Effect of two simultaneously contracting centers on a common meshwork.

expand, as happens, for instance, in a vesicle or tube or cyst swelling by the secretion of liquid into its closed lumen. Evidently, as shown in Fig. 12, top, the meshes of the fibrous medium become circumferentially compressed so as to assume a predominantly tangential orientation around the fragment. Cells happening upon such a territory would obviously be forced to circle round and round producing an envelope or tunic. Many of the connective-tissue sheaths and capsules in the meshwork of the body owe their origin to this mechanism. By contrast, if somewhere in the meshwork there arises an area which contracts, the fibrous components will be gathered toward that center in a predominantly radial orientation (Fig. 12, bottom). Cells in such zones then become likewise disposed radially. This type of effect obtains frequently in the vicinity of rapidly proliferating cell groups. It seems that chemical agents, as yet undefined, are being discharged by such cells which cause intensive synaeresis of the surrounding colloids, which means condensation of the fibrous components with loss of bound water, the resulting local shrinkage being much greater than the gain of mass by cell growth. In other words, a purely *scalar* change in a piece of tissue—increase or decrease of volume—can, through the intermediary of a fibrous continuum, translate itself into marked *vectorial* effects, establishing well-defined geometrical and structural patterns. A further degree of ordered complexity is introduced if two or more cell masses are explanted together in a common clot (Fig. 13). Through their joint constriction effects, they deflect the fibrous meshes of the intervening medium into a line connecting the active centers, and this, of course, constitutes a path for direct cell traffic between them. The cells growing out from the two centers simply follow the submicroscopic bridge which has been laid down for them automatically by the tension-engendering chemical activity of their sources (Fig. 14). Here is one primitive example of how chemical action can translate itself into physical organization.

Extrapolating briefly from these model systems of living cells, one may assume that the same sort of intimate interdependence shown here for the microscopic dimension repeats itself both in submicroscopic and in higher supracellular dimensions. One is led to the conclusion that there is a tie between physical structure and chemical activity which is indissoluble and which

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

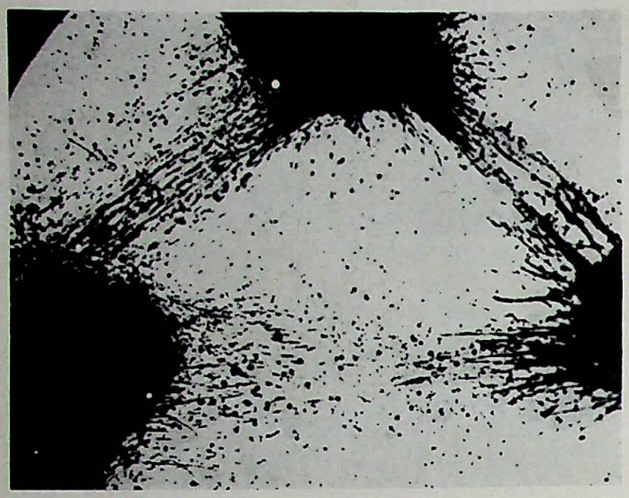


FIG. 14. Automatically established triangular cell-bridge connections between three separate tissue cultures (dark masses) in a common blood-plasma membrane.

it is not enough to assert, but which has to be explored systematically. Where, in order to study a particular metabolic reaction in isolation, the biochemist provides optimum conditions for that process, the cell produces the unique prerequisites for that same reaction only in certain strictly confined localities. It does it through the metabolic products of other chemical reactions going on in another equally confined sample of cell space. Thus, one system feeds another and, in turn, is fed by a third, in a vast system of mutually interdependent, symbiotic, and harmonized partial reactions subsidiary one to another. This interdependence, characteristic of the living state, is what I once termed "molecular ecology." As a field of investigation, it barely has reached infancy. Yet its significance is pointed up by the fact that, in the living cell, the various biochemical partial systems coexist in a common space, without preformed rigid partitions, but rather compartmentalize themselves by structural effects of their own activities of the sort exemplified in crude and elementary fashion by the case just cited. Physical structure and physicochemical conditions limit the types of enzymatic and other chemical reactions that can go on in a given spot —for instance, along interfaces of fibrillar, lamellar, or corpuscular systems—while the resulting reactions, in turn, modify the physical substratum; and by continual interplay of this sort between physical structure and chemical action, the cell system passes first through its progressive developmental transformations and then is stabilized in the steady state of maturity. To some extent, physical structure then is frozen into static arrangements of cytoskeletons, but even then physical structures still are regenerated continuously by cellular activity, leaving at least part of the cellular system in a state of incessant development and self-renewal.

A realistic concept of cell behavior also must take into account the limits set to interactions between distant parts by the formation of compartments within compartments. A diagram of the organism (Fig. 15) would represent it as a system of concentric shells, with the gene in the center enclosed in the chromosome, which is enclosed in the nucleus, which is enclosed in the cytoplasm, which forms part of a tissue, which forms part of the organism, which is surrounded by the external environment; for simplicity, cytoplasmic particles are omitted. No outer agent can influence any of the inner shells except through the mediation of the shells in between, which may or may not modify that factor during its inward passage. Conversely, products of inner systems may not reach outer shells as such, but may be significantly screened and altered in transit. The arrows in the diagram indicate the complex network of relations that one must bear in mind. Oversimplified mental pictures form when, for instance, the simple statement is heard that a gene "controls" a particular feature of the organism, without allowing that it can do so only through interactions with the outer shells, which in themselves have become progressively modified in their long developmental history by countless chains of interactions with other shells, including, of course, the innermost, the gene. Since it has been my assignment to sketch the living cell as it truly is in all its highly ordered complexity, I feel compelled to caution against the illusion that a simple statement, such as, "a gene controls a character" reflects any similar degree of simplicity in the phenomenon covered by the statement.

A final example projects the principle of interaction between physical structure and chemical activity upward into the realm of supracellular order, a field so baffling in its problems that many an investigator prefers to look the other way when he encounters them. The example is chosen from an almost diagrammatically simple object which, because of this, holds at least some promise of more penetrating analysis by the combined physical and chemical tools now at one's disposal. It refers to the origin of the internal architecture of cartilage. Cartilage is formed by groups of cells producing

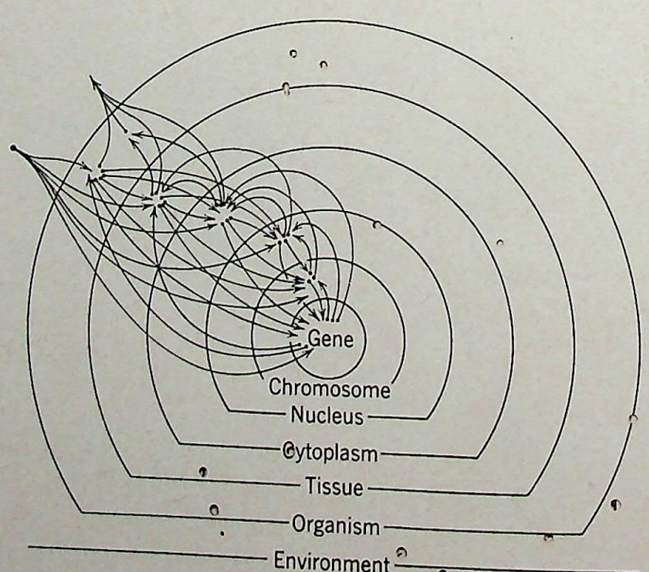


FIG. 15. Diagram of network of possible interactions in an organism.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

ground substance between them. That ground substance has been identified partly as a mucopolysaccharide—chondroitin sulphuric acid—and collagen. But cartilage is not just chemical substances; it assumes characteristic shapes and configurations, depending upon its sites, owing to developmental processes under genic control. The nature of the problem becomes clear when one contemplates that the peculiar pattern of convolutions of one's earlobes, which are but covered cartilage, is an individual inherited characteristic. How do such patterns form? Cartilage of the limb is different from cartilage around the eye; the former grows in compact, whorl-shaped masses, the latter develops as a flat plate. It has long been known that, if the cell group that is destined to form limb cartilage is reared in tissue culture, it will grow according to the standard limb pattern, giving rise to recognizable skeletal elements. Similarly, the author observed years ago (jointly with R. Amprino) that the precursor cells of the scleral cartilage around the eye, after explantation as an intact group, go on to form cartilage in the shape of a plate. Evidently, in either case, the explanted tissue complex contained some physical properties that guided the cells contained in it into their respective typical arrangements and growth patterns. The crucial property differentiating between limb and eye cartilage had to be thought of as inherent in the block of tissue as a whole, the further development *in vitro* merely amplifying some distinctive architectural pattern already present in the tissue fragments at the time of their isolation from the embryo.

To test this conjecture, the author (with A. Moscona) recently resorted to a technique that permits one to disintegrate a tissue into its constituent cells and thereby to destroy any intercellular structures and supracellular arrangements that could have acted as guides for subsequent development. Cells can be separated from one another and from their matrix by trypsin, washed in a loose suspension, and then seeded out in a tissue culture where they can aggregate into random clusters. More about the manner of aggregation is reported in another article (Weiss, p. 449). When a piece of prospective limb cartilage and a piece of prospective eye cartilage were dissociated in this manner into their component cells and the cells of each type were permitted then to aggregate in tissue culture and to continue with their actual cartilage-forming activities, it turned out that the cells from the limb site produced cartilage of the typical whorl-shaped massive limb pattern (Fig. 16, top), while the cells that had originated at the eye site produced the plate-shaped laminated structure to which they would have given rise prior to their disaggregation (Fig. 16, bottom). In other words, the "blueprint" of the architecture of the group performance is ingrained in each individual cell and not just carried over into the culture by the block of tissue as a whole. Now, how to convert this figure of speech into concrete terms is problematical. But one may speculate that the architecture lies in the specific

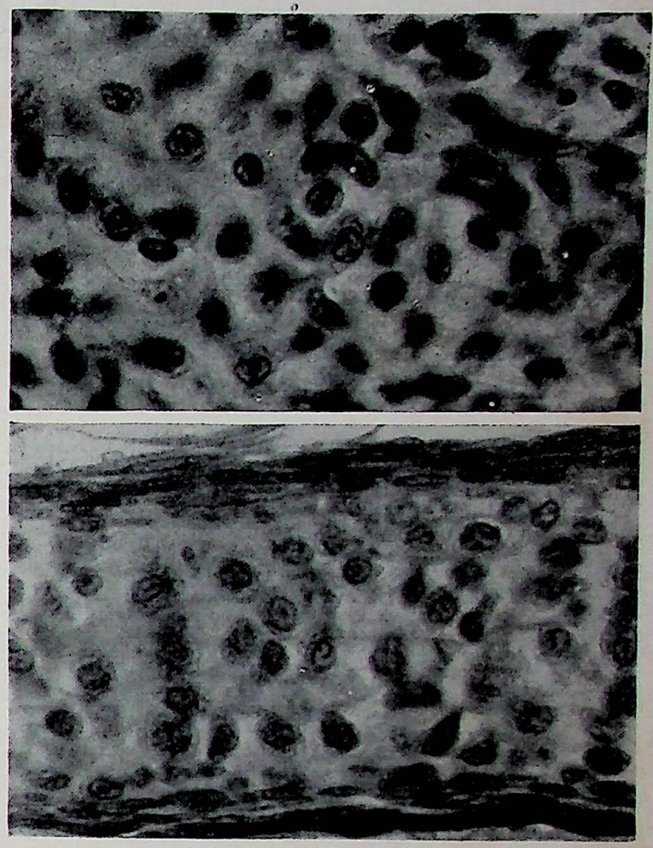


FIG. 16. Cross sections through cartilages developed in tissue culture from dispersed and reaggregated precartilage cells from a prospective limb (top) and from a prospective eye coat (bottom) of the chick embryo, forming whorl-shaped and lamellar structures, respectively.

ground substances secreted by the cells. One would have to suppose that each cell type secretes a complex ground substance of a distinctive pattern of macromolecular stacking or crystallinity of such a kind that it would predispose a planar array of the mass in the form of layers in one case, and a more massive isotropic arrangement in the form of three-dimensional whorls in the other. The cells then would dispose themselves in conformity with these structural patterns of ground substance for which they had furnished the elements in the first place.

This is sheer conjecture. No facts are known that would either support or contradict it. Yet this very uncertainty helps to point up the immensity of ignorance in matters of supracellular organization. If one is convinced that such higher-order organization is to be explained solely in terms of cellular dynamics, then one must raise one's sights to the domain where those phenomena occur, which is no longer the intracellular microcosm.

The study of tissue architecture, used here as an example, may be one of the easier inroads into the maze of perplexing problems presented by the orderliness of the complex living organism, as against the relative disorder and simplicity of its shattered fragments, on which one preferably concentrates, mostly after homogenizing them either physically or conceptually.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

The lessons of this story repeat themselves as the viewpoint is shifted from the organization of the body in its structural aspects to the supracellular coordination and ordering systems of its functional activity, whether this be the homeostatic maintenance of blood composition, the integrative action of the nervous system in behavior, or the mobilization of defense and repair mechanisms in response to pathological disturbances. Of course, the operative tools in all of these performances are the individual component cells, which in turn operate through physically ordered subsystems in incessant chemical interaction. But unless one remains cognizant of the fact that the level of the organism is reached from the level of molecular biology not in one single jump over the conceptual gap customarily bridged by the word "organization," and unless one learns to think in terms of a hierarchy of ordered systems, each one with some degree of identity and stability even on supramolecular levels, one obscures rather than enucleates the problems to which thinking and research must be directed. The most impressive feature of a cell is not the constant flux, reshuffling, and variability of its population of molecules and particles (of which our cinematographs presented at the Study Program have given a clear expression), but the fact that, in spite of this ever-present change, each cell remains so remarkably invariant in its total behavior; that indeed, as an entity each behaves so much like millions of other entities equally variable in inner detail, that one comes to recognize them as essentially alike. Such relative invariance of the whole presupposes the harmonious subordination of the behavior of the parts to the conditions of the collective group. It presupposes that the free interactions among the subunits are subject to restraints, the nature and direction of which vary adaptively with the state of the system as a whole. I can think of no more propitious introduction to a biophysics program than by re-emphasizing that the restraints in question are not the sole province of chemistry, especially the stoichiometric branch, but that they include crucial restraints of *physical* nature, with nonrandom molecular arrays a most prominent and analytically most promising feature.

Awareness of the existence of these problems in no way detracts from the pragmatic value of analyzing each elementary component process in its own right and of using whatever models, however simplified, may be found to be constructive as aids to understanding. But it is precisely the phenomenal advances made in the study of the partial and isolated system, in contrast to the dearth of information on the *organized interactions of such systems in the living cell and organism*, that lead me to conclude with a plea for a more balanced effort and more extensive occupation with the latter problems than there are at present. Cellular and tissue dynamics are population dynamics of species of molecules, of cellular subunits, and of cells. Such group dynamics cannot be derived solely from looking at the members of the population in isolation, but only from a study of the ecology and technology of their behavior in the group. It is hoped that this modest effort at presenting the case of the living cell not only has given a fair introduction to the problem but also has shown that practical techniques for its eventual solution are at hand, with physical and chemical approaches indissolubly interwoven.

BIBLIOGRAPHY

P. Weiss in *Chemistry and Physiology of Growth*, A. Parpart, editor (Princeton University Press, Princeton, New Jersey, 1949), p. 135.

P. Weiss and B. Garber, Proc. Natl. Acad Sci. U. S. **38**, 264 (1952).

P. Weiss, J. Embryol. Exptl. Morphol. **1**, 181 (1953).

P. Weiss in *Analysis of Development*, B. H. Willier, P. Weiss, and V. Hamburger, editors (Saunders Company, Philadelphia, 1955), p. 346.

P. Weiss, J. Cellular Comp. Physiol. **49**, Suppl. 1, 105 (1957).

P. Weiss, Intern. Rev. Cytol. **7**, 391 (1958).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

REVIEWS OF MODERN PHYSICS VOLUME 31, NUMBER 1 JANUARY 1959

R 530.5 AM3R V.31]
112457
Gurukul Kangri Vishwavidyalaya 596 Library

3
Cellular Responses

WILLIAM BLOOM
Committee on Biophysics, The University of Chicago, Chicago 37, Illinois

THIS paper is concerned with some aspects of the organization of cells, particularly as seen with the light microscope, and the results of some experiments on dividing cells in tissue culture as shown by time-lapse motion pictures. The paper is illustrated by portions of a few key frames from the motion pictures.

There have been several important periods in the study of cellular organization. One of them was the period of descriptive histology, embryology, and cellular pathology of the nineteenth century. In it a variety of important processes in the development and function of cells were portrayed largely, but not exclusively, on the basis of dead material. This was followed by a period of extreme skepticism in which a number of investigators claimed that nearly everything seen after fixation and staining was artifact; they believed that only what could be seen in the living cell existed in it. There were even claims that nuclei could not be seen in living cells and hence were artifact. Actually, nuclei are easily demonstrable during life. This period of skepticism was helpful in some respects but it also did great harm because some of its claims were too destructively critical. One of the rewards of living in the present period is to see that many of the so-called artifacts of the previous century have been shown to exist in living cells by our more modern methods of investigation. I remember distinctly the first time I saw a living cell with the newly developed phase-contrast microscope. What I saw was like an iron-hematoxylin stained cell moving around under the microscope and offering complete verification, at the light-microscope level, of the existence of certain structures previously known mainly in preparations of dead cells.

As our methods of morphologic, chemical, and physical analysis become more powerful and as the study of the minute structure of living systems reaches ever closer to the molecular level, it becomes more and more obvious that the old problem of what is artifact in our observational material is still with us.

I shall show a series of newt cells cultured in diluted chicken plasma. In the time-lapse movies, constituents of a typical cell are shown as it goes through the process of cell division (mitosis). In the film, the reality becomes apparent of some structures that to some of you may have been only words. Three cells show the effects of either ionizing or ultraviolet, total cell irradiation. The succeeding five dividing cells demonstrate some of the effects of localized irradiation of small parts of them. In the cells which were irradiated throughout their extent, several abnormalities appear, while in those irradiated locally, the number of abnormalities is much smaller. The comparison of local *versus* total cell irradiation gives a little insight into some aspects of the machinery of cell division, a problem which Raymond Zirkle and the author began to work on some eight or nine years ago with the aid of our proton microbeam.[1] A few years later, Robert Uretz began to work with us in our study of mitosis; the simple ultraviolet microbeam, which he devised, has been of great help.[2,3] With it, one can aim directly at a structure and watch it as it is being irradiated.[4,5] Others in our laboratory, especially E. W. Taylor and R. P. Perry, have also contributed to our knowledge of what happens in dividing cells. Zirkle[6] has recently reviewed the methods and results of microbeam irradiation which started with the work of Tschachotin with ultraviolet light in 1912.[7]

In our proton microbeam, the protons start with an energy of 2.0 Mev but lose perhaps a quarter of this as they pass through two layers of mica, each 5 μ thick. One of these seals the delivery opening of the Van de Graaff generator; the other is the cover slip for the culture. Although there is a small percentage of scatter, if the culture is essentially in contact with the microaperture, 85 to 90% of the protons bombard an area about 2.5 μ in diameter, and practically all of the protons pass within a circle of 8 to 10 μ diameter at the target. (The indicated variations reflect slight differences in approximation of the mica cover slip to the microaperture.) The scatter is not important with small numbers, but obviously becomes serious with hundreds of thousands. With a more powerful generator, it would have been possible to deliver a single proton into a target of 1 μ^2, but the smallest number we have been able to deliver with our generator is a burst of about 20.

The Uretz microbeam makes use of a reflecting objective which brings all wavelengths of visible and ultraviolet light into approximately the same plane. This device has an important advantage over the proton microbeam in that the size and shape of the area to be irradiated can be varied easily from circles 2 μ in diameter to much larger ones, or to slits, or to irregular shapes of various dimensions. We can also compare the effects of microbeams of monochromatic and heterochromatic ultraviolet light with those of the proton microbeam.

The machinery by which a plant or animal cell becomes two cells is probably much the same in all, although much more is known about the process in nucleated cells than in bacteria. Much of what is shown here has been known in the descriptive sense since about 1880 in the work of Strassburger and especially of Flemming, among others. But we know very little more today than these early workers knew about the causal

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

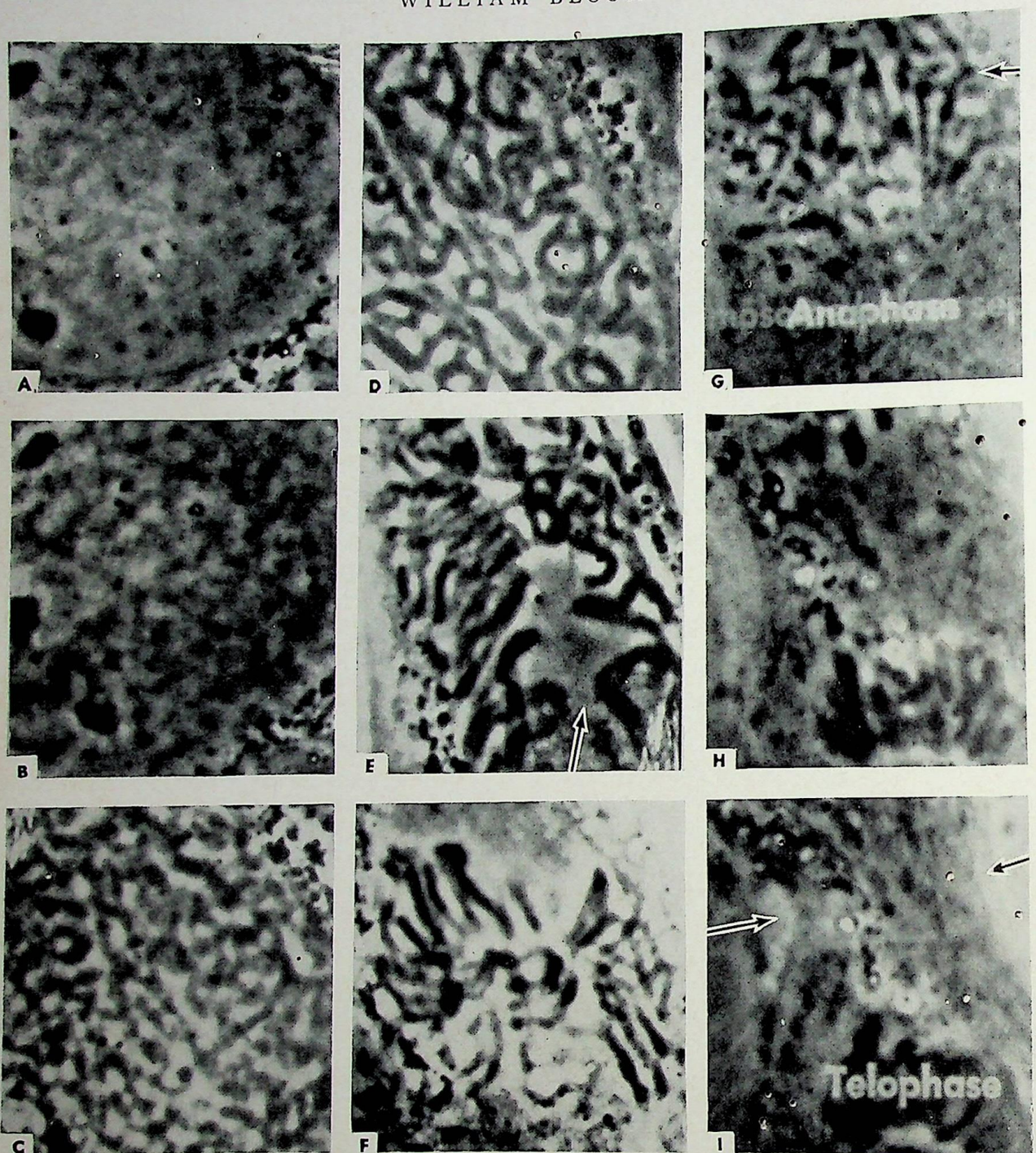


FIG. 1. This sequence [(A) to (I)] shows some of the main stages in mitosis. The nucleus at (A) is in a very early stage in the change of a relatively nongranular nucleus into chromosomes. The two large bodies are nucleoli. At (B), the cell has moved slightly and three nucleoli are visible. The nuclear granules begin to form vague cords which are much more definite at (C) and especially so at (D) where they have become typical chromosomes. (E) shows the change in arrangement of chromosomes from prophase into metaphase at (F). In (G), each chromosome is dividing into two daughter chromosomes (as at arrow) which are beginning to move to opposite poles of the cell (anaphase). At (H), the chromosomes have completely separated and in (I) reconstruction of the daughter nucleoli is beginning and the cell body is constricting (arrows). The arrow in (E) points to a centriole.

mechanisms involved in this complicated, meticulous process by which the two daughter cells inherit equal amounts of genetic material.

In the films, the area of the cells shown is approximately 40 by 60 μ, and the motion pictures are speeded up about 100 times. When present, the clock indicates the lapse of time. Figures 1 to 9 are from the time-lapse motion picture shown at the presentation of the paper, photographed with medium dark, phase-contrast microscopy; 1400×. All cells are from newt heart explanted in a mixture of one part chicken plasma and nine parts amphibian Tyrode solution. The division process in these cells at 70°F may take as long as 6 or 7 hr or as little as 4 hr. These differences, at the same temperature, are unexplained.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

UNTREATED CELL IN MITOSIS

In cell 1, a typical mitosis, one can see tiny fat droplets surrounding the nuclear membrane, the filamentous and granular mitochondria, the nucleus with large nucleoli, and numerous, irregular small granules [Fig. 1(A)]. The author used to question whether the movement of granules in the cytoplasm was Brownian motion, since one sees it only in the speeded-up time-lapse movies, and not by direct observation of the living cells. But Taylor studied this phenomenon and convinced him that it was Brownian motion.

As the film continues, the small granules in the nucleus become more numerous and darker and gradually form long snake-like bodies—the chromosomes [Figs. 1(A)–1(D)]. They are about 2 μ thick and some 30 μ or more long. The nucleoli disappear suddenly and we do not see what happened to them [between Figs. 1(B) and 1(D)]. The nuclear membrane also disappears and about the time that it does the chromosomes begin a continuous, irregular movement. Warren Lewis, who has spent many years studying living cells, called this the "Dance of the Chromosomes."

The cell contains two other important small structures called centrioles, one above and another below the group of chromosomes. The poles of the spindle are connected with the centrioles. The area with practically no granules is part of the developing spindle [Fig. 1(E)]. After some time, a pale stripe appears within each chromosome and extends through its length. Suddenly, the chromosomes divide along this stripe and the daughter chromosomes (or chromatids), each a half of an original chromosome, move apart [Fig. 1(G)] and form the two separate daughter nuclei [Figs. 1(H) and 1(I)]. Then the cell body constricts, the whole mass separating into two new cells, each with a nucleus. The temporary excrescences on the surfaces of the cell are normal when they are not more extensive than in this cell. In the irradiated cells, which are seen next, they may be much larger, more numerous, more violent in movement, and have longer persistence.

The process of cell division goes on in our bodies hundreds of millions of times a day in order to replace those cells which are normally lost for various reasons. In the vast majority of cases, these mitoses are executed as meticulously as in this cell.

RESPONSES TO TOTAL CELL IRRADIATION

The next three cells show a number of different phenomena as a result of total cell irradiation. Cell 2 received 200 r of 200-kv x-rays. The film starts with the cell in an early stage of mitosis. As a consequence of the irradiation, an abnormality develops instead of the orderly cell division which occurred in the untreated cell. Chromosomes start to separate but the separation is not complete [Fig. 2(B)]. Then cell constriction occurs about the mass of chromosomes [Fig. 2(C)] and forms a dumbbell-shaped nucleus connected by a long bridge of chromosomal material extending between the two daughter cells [Fig. 2(D)].

Cell 3 received 350 r of 200-kv x-rays. Instead of producing a relatively simple malformation, as produced by 200 r in the previous cell, several abnormalities occur. The chromosomes make an abortive attempt to separate into their daughter chromosomes [Fig. 3(A)], violent "bubbling" appears on the cell surface, and two constrictions of cytoplasm occur [Fig. 3(B)]. We have seen bubbling go on for days in cytoplasm pinched off as in this case. Nuclear material now separates into two irregularly sized groups [Fig. 3(C)], but at one point there is a tiny connection between them, and the smaller group of chromosomes spurts over into the larger one, making a bi-lobed nucleus.

Cell 4 received ultraviolet light over its whole body. Despite the marked abnormalities which result, the cell will eventually give rise to daughter cells which are able to exist for a time at least in tissue culture. Immediately after being irradiated, some of the motion of structures within the cell ceases, but some of it gradually returns. For a short period, there is an irregular arrangement of chromosomes and then terrific bubbling movements begin to appear [Fig. 4(B)]. The chromosomes and cytoplasmic granules move in and out of the bubbles. We call this "type 4 bubbling," the most extreme degree. The term "bubbles" is a misnomer as they are protuberances of cytoplasm and not bubbles; their nature is completely unknown. Through it all, the chromosomes stick together, more or less. The bubbling motion gradually begins to subside. Then the chromosomes separate into one large and one small group which move apart and a constriction of the cytoplasm develops [Fig. 4(C)]. As nuclear reconstruction begins, it becomes apparent that the two nuclear masses are connected by a thin chromosomal bridge.

RESPONSES TO PARTIAL CELL IRRADIATION

In the next 5 cells, very small areas of nucleus or cytoplasm were irradiated by protons or ultraviolet light.

There is a widely held view in the radiobiological literature that chromosomes are damaged secondarily as a result of irradiation of the cytoplasm. The experiments on the next two cells clearly refute this claim for the newt cells in our culture. In cell 5, the area indicated by cross-hairs [Fig. 5(A)], about 14 μ from the nearest chromosomes (arrows), was irradiated with 28 300 protons. The intensity of irradiation of this bit of the cell is enormous when one realizes that each proton equals 600 rep. Nevertheless, the resulting mitosis is perfectly normal [Figs. 5(B) and 5(C)].

Cell 6 received 60 protons localized over a small area of two chromosomes [see cross-hairs and arrowhead in Fig. 6(A)]. The daughter chromosomes on the nonirradiated side separate cleanly, while those on the irradiated side are stuck together. The daughter chromo-

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

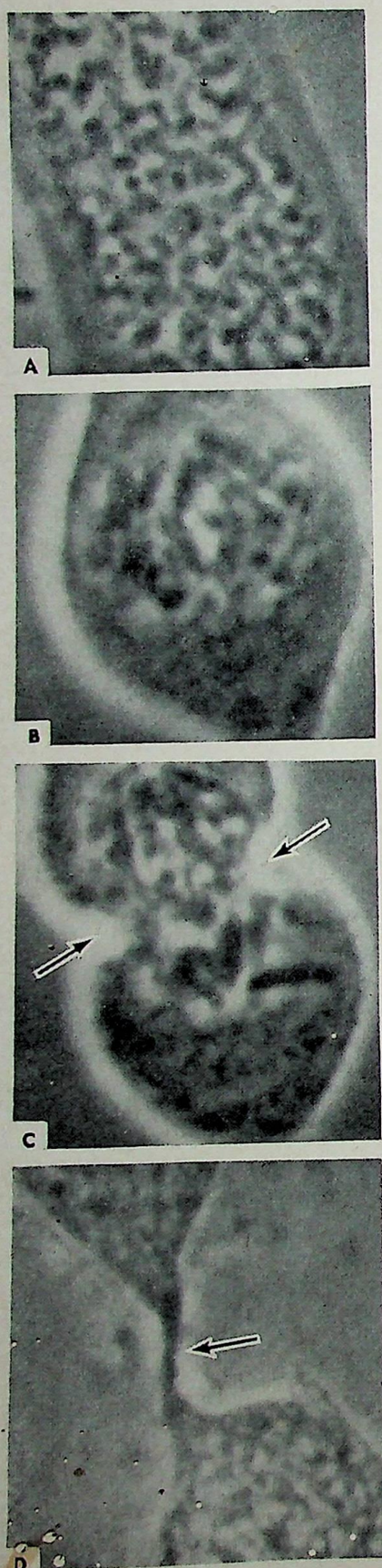


FIG. 2.

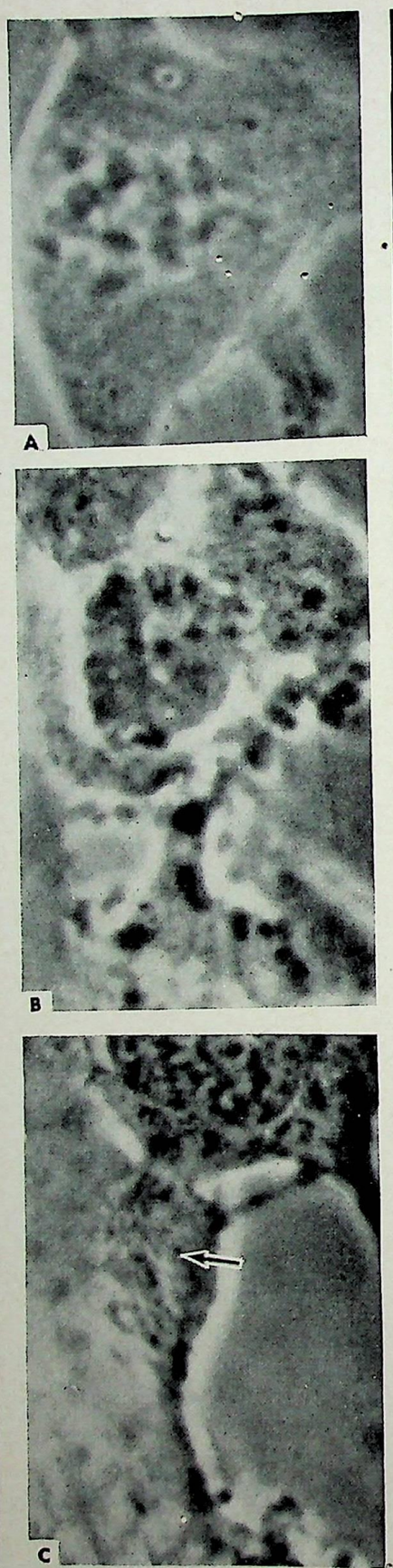


FIG. 3.

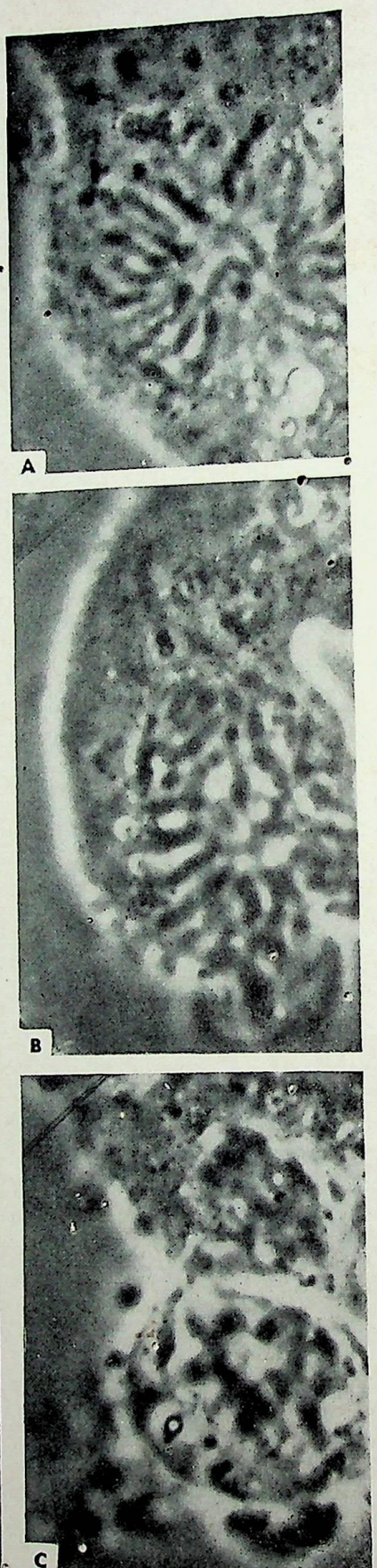


FIG. 4.

FIG. 2. This cell received 200 r of 200-kv x-rays. (A) is immediately after irradiation and shows no change. (B) was taken during the abortive anaphase. This is the maximum that the daughter chromosomes have separated. At (C), the constriction of the cell (arrows) includes the chromosomal mass and produces a dumbbell-shaped nucleus (D) with a thin bridge (arrow) connecting two main portions of the reconstructing nuclear material. (*Note.*—Captions for Figs. 3 and 4 are on page 25.)

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

somes move apart as a group, much as in an opening hinge pinioned on the irradiated parts of the chromosomes [Fig. 5(B)]. This produces a temporary chromosomal bridge [Fig. 6(C)]. The difference between the failure of the 28 300 protons delivered to cytoplasm to produce any visible change in the chromosomes of the previous cell is in sharp contrast to the effect of 60 protons delivered locally to chromosomes in this cell.

Areas 8 μ in diameter of the next three cells were irradiated with ultraviolet light. The first of these cells (number 7) is in early mitosis (prophase). The spindle has not yet formed and, as a result of cytoplasmic irradiation, will not form. There is an irregular arrangement of chromosomes characteristic of this stage. After the localized, cytoplasmic, ultraviolet irradiation, the irregular configuration persists [Fig. 7(B)]. The chromosomes shift around from time to time and there are pulsations and temporary extrusions of the cell surface. After several hours, a regrouping of the chromosomes occurs suddenly. They do not give rise by longitudinal splitting to a double number of daughter chromosomes, as in normal mitosis. Instead, haphazard numbers of *whole* chromosomes aggregate into two groups which separate one from another [Fig. 7(C), arrows]. Then the cytoplasm constricts and two daughter cells are formed. Such cells obviously differ completely in their genetic complements from each other and from those cells arising from a normal mitosis.

In the next cell (number 8), it is possible to follow the aggregation of the two groups of chromosomes more easily than in the previous one wherein cytoplasmic irradiation had prevented the appearance of the spindle. In this cell, which is in the mid-phase of mitosis, the spindle is fully developed. The area to be irradiated is shown at the cross-hairs [Fig. 8(A)]. After irradiation of a small part of the cytoplasm (8 μ) with ultraviolet light, the spindle breaks down and the chromosomes lose their characteristic, butterfly-like arrangement, becoming irregularly disposed in the center of the cell [Figs. 8(B) and 10]. A pale longitudinal line becomes visible in each chromosome and marks the usual line of cleavage of each chromosome into its two daughter chromosomes (chromatids). The pale line disappears, however, before the whole chromosomes aggregate into two irregular groups [Fig. 8(C), arrows]. In cells so treated, the number of chromosomes in each new nucleus may vary from 9 to 15 or even more, instead of 22 which is characteristic of this species. Constriction occurs and the daughter cells move apart.

Since, in the experiments on cells 7 and 8 (and nearly 100 others similarly treated), irregularly sized groups of whole chromosomes move apart in the absence of the spindle, it would seem that the spindle serves primarily as a mechanism by which *daughter chromosomes* are guided as they are moved apart by unknown forces.

From the days of Flemming on, there have been many theories about the machinery of mitosis. Some of the models, such as those based on rubber bands, have some resemblance to what is seen in some stages of mitosis, but they contribute nothing to an understanding of the forces determining the processes in mitosis. It has been claimed that the daughter chromosomes become separated by electrostatic forces or by a localized imbibition of water, or that the spindle is made of a contractile material which pulls them apart. Although cognizant of many claims to the contrary, it is probably fair to say that there has been no delineation at all of the nature or amount of the forces involved in mitosis.

Nor do we have any idea of the nature of the force which moves the irregular groups of whole chromosomes apart in the absence of the spindle. It is known that, in the early stages of mitosis, the two centrioles separate and take position at the apices of the spindle, and that there is a temporary attraction between the centrioles and the kinetochores of some of the chromosomes. Then, this attraction ceases and all the chromosomes move into a position midway between the two centrioles (metaphase). In the next stage of division (anaphase), the daughter chromosomes separate from their mates and move with their kinetochores foremost toward the centrioles. In those cells in which the spindle is abolished, we assume that, during the separation of the groups of whole chromosomes, those chromosomes close to each of the centrioles go with it to a daughter cell.

The last cell (number 9) shows a different phenomenon after localized irradiation (8 μ in diameter) of chromosomes with ultraviolet light [Fig. 9(A)]. A dramatic change in the refractive index of the chromosomes appears in the irradiated area and produces what we call a "pale" spot [Fig. 9(B), arrows]. This change does not appear if much more than half of the chromosome mass is irradiated. However, irradiation of as little as 2 μ of a chromosome will produce it. The mass of chromosomes immediately around the pale area becomes sticky and cannot separate into chromatids, while those some distance away can. From this there results an irregular, lobated single nucleus with malformed chromosomes and a pale spot in the center [Fig. 9(C)].

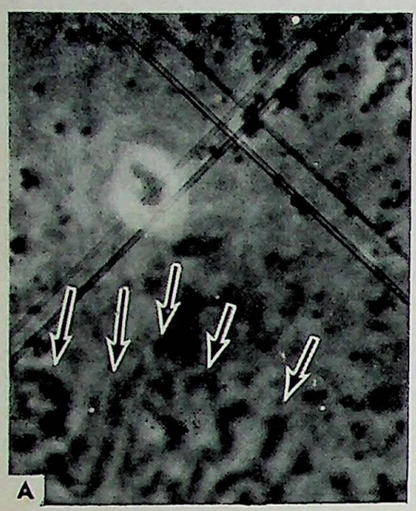
In fixed preparations, the pale area also appears pale with most of the usual stains. A similar appearance

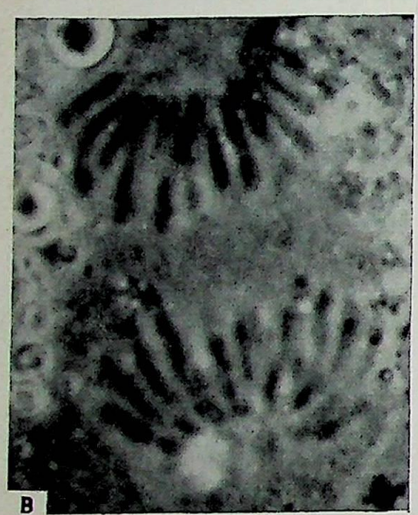
---

FIG. 3. This cell was irradiated with 350 r of 200-kv x-rays. In it the anaphase movement did not progress as much as in the previous one [compare with Fig. 2(B)]. Shortly after (A), the cell begins to show violent "bubbling" and multiple constrictions. As a result, portions of cytoplasm with or without chromosomal material become separated. At (C), the uppermost part of the picture shows a newly formed daughter nucleus and attached below it is an irregular chromosomal mass (arrow).

FIG. 4. This cell received a 2.5-min exposure of a germicidal ultraviolet lamp 2.5 cm from the culture. At the time of irradiation, the cell showed typical metaphase configuration. At (A), the chromosomes show an irregular rosette appearance. A few minutes later, the cell begins to undergo violent bubbling and at (B) the chromosomal mass is being extruded into one of the bubbles. Later, there are multiple constrictions (C).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA





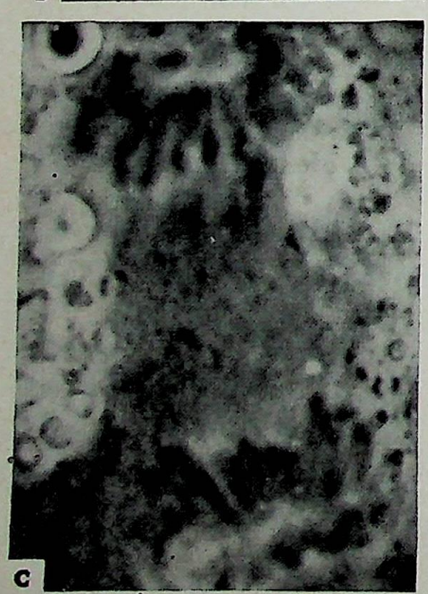
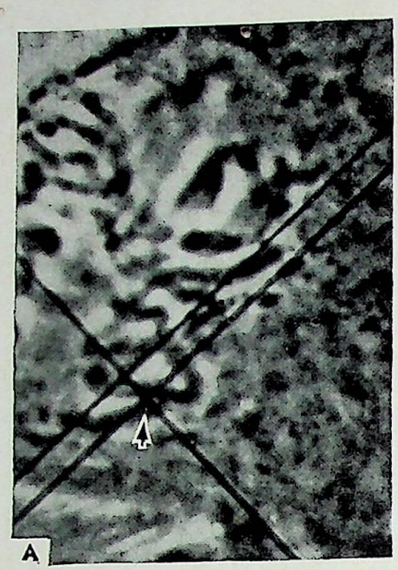
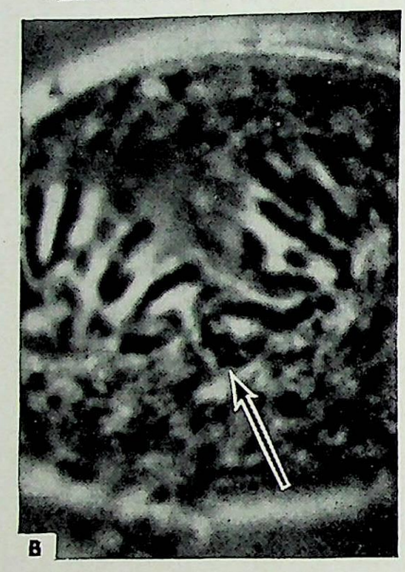


FIG. 5.





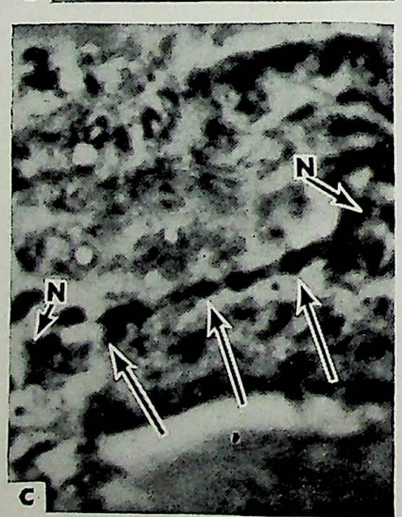


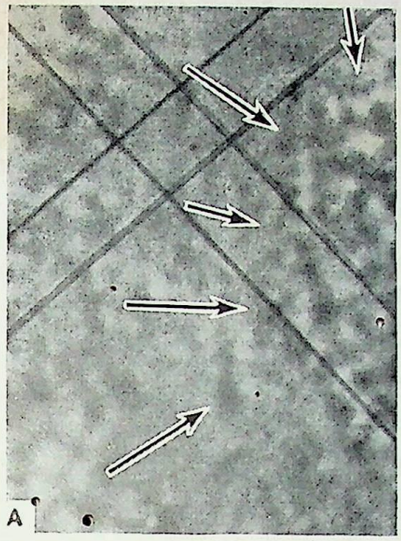
FIG. 6.

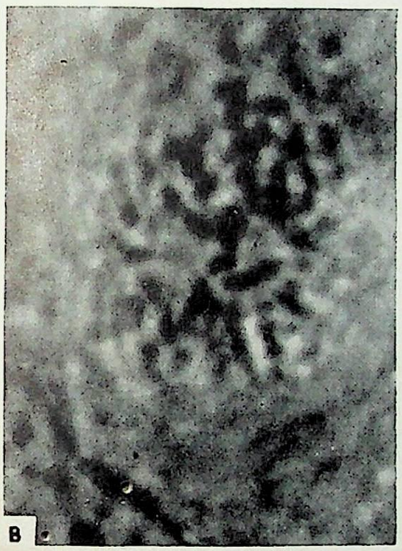
FIGS. 5 AND 6 show the difference in effect of localized irradiation of cytoplasm *versus* localized irradiation of chromosomes in dividing cells. The cell shown in Fig. 5 received 28 300 protons at the area indicated by the double cross-hairs. The nearest chromosomes are indicated by the arrows. This cell divided into two by a perfectly normal mitosis as shown in (B) and (C). Figure 6 shows a totally different picture as a result of irradiating small parts of two chromosomes (indicated by cross-hairs and arrow) with 60 protons during metaphase. As shown in Fig. 6(B), the chromosomes separated normally into daughter chromatids except those at the point of irradiation where, as indicated by the arrow, the chromosomes are stuck together. As a result, the two daughter nuclei (N,N) are joined by a chromosomal bridge indicated by the three arrows.

results from the application of the Feulgen stain for deoxyribonucleic acid.[8] This suggests that, in the development of the pale spot, the nucleic-acid moiety of the irradiated part of the chromosomes may have been so altered that it does not respond to the stain or has actually left the area. Ultraviolet-absorption studies of the pale spot by Perry show a diminished absorption at 260 mμ (Fig. 11).[8]

Zirkle has spent much time in the past few years determining the relative effectiveness of the following wavelengths (225, 240, 250, 260, 270, 280, and 300 mμ) on paling. He found the action spectrum to have the

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA





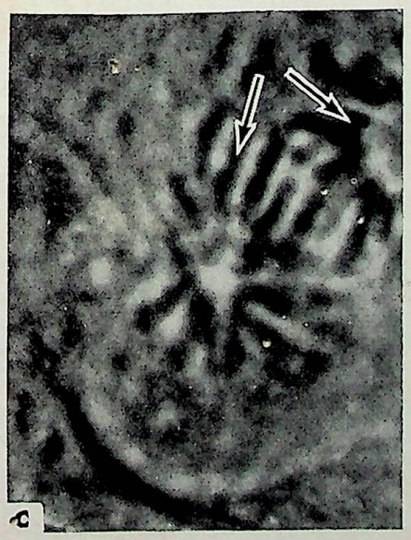
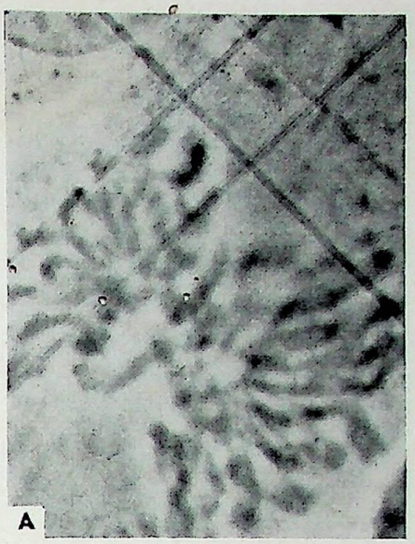
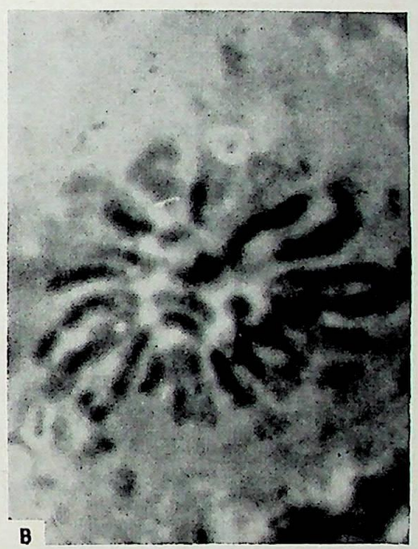


FIG. 7.





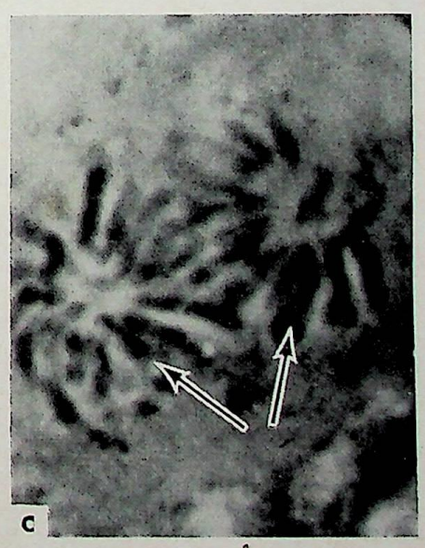
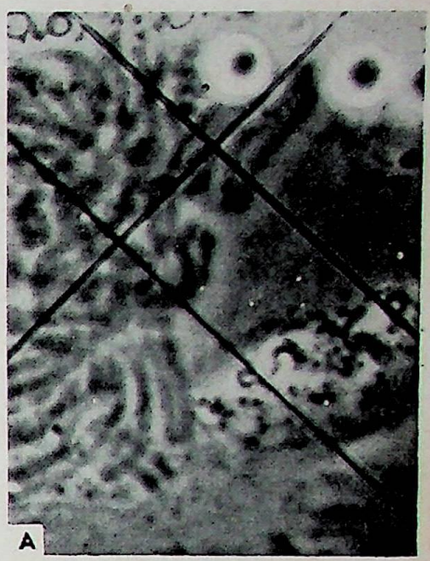
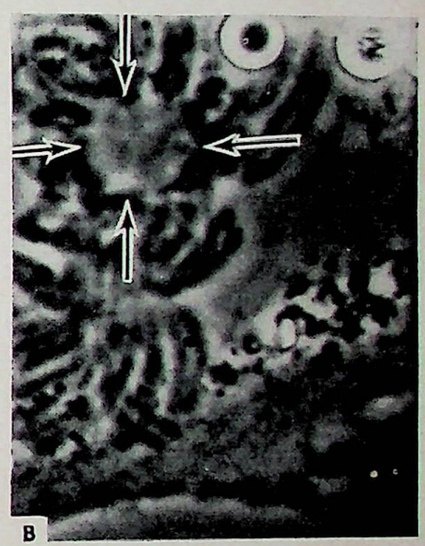


FIG. 8.





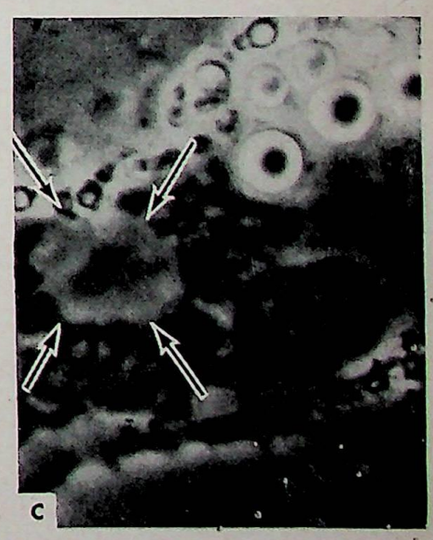


FIG. 9.

FIG. 7. This cell in prophase received localized, ultraviolet irradiation at the area indicated by the cross-hairs. The nearest chromosomes are shown by the arrows. The spindle did not form and the chromosomes occupied an irregular clumped area (B) and later separated as whole chromosomes into two unequal groups, indicated by the arrows in (C).

FIG. 8. The cell, in metaphase with a clearly marked spindle, received localized, ultraviolet irradiation of the area indicated by the cross-hairs. The spindle disappeared promptly and the chromosomes show a haphazard clumping (B) much like that in 7(B). In cell 8, as in the previous one, whole chromosomes separate into two daughter clumps, shown by the arrows in (C).

FIG. 9. This cell, in metaphase, received ultraviolet, localized irradiation of an area of chromosomes 8 μ in diameter, indicated by the cross-hairs. Within a minute, the chromosomes showed a typical "paling" reaction (B). The cell attempted to divide but was unable to execute a complete anaphase as the chromosomes immediately around the pale area were very sticky. There resulted a single irregularly lobed nucleus with a pale spot (arrows) in (C).

CC-0. Panjab University Haridwar Collection. Digitized by S3 Foundation USA

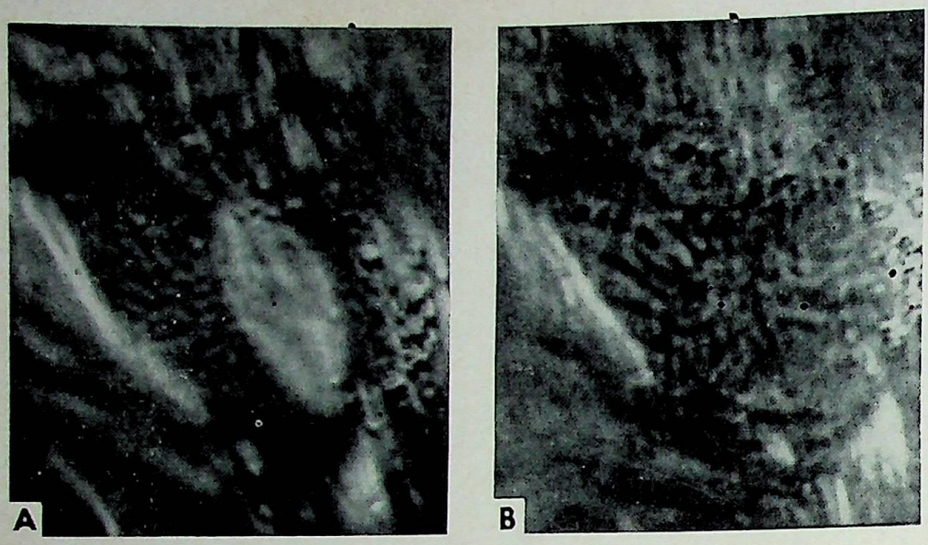
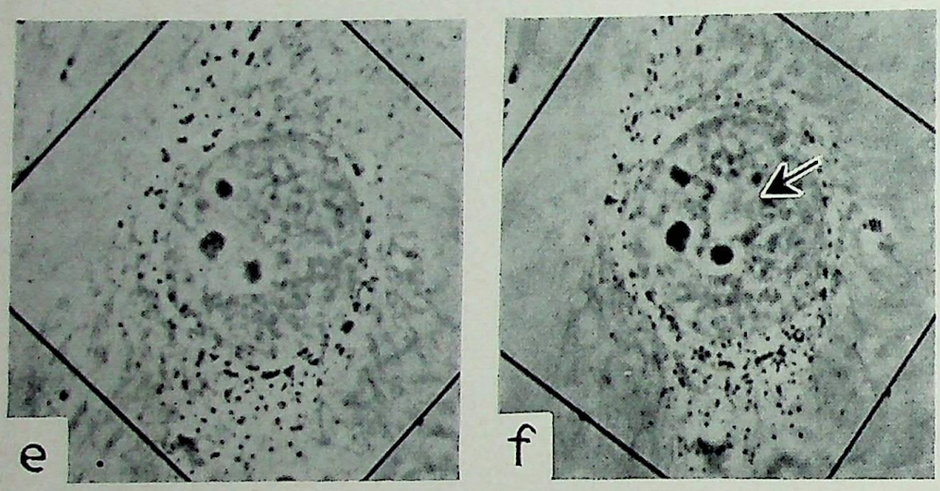


FIG. 10. Newt cell in metaphase as seen with polarized light. (A) The spindle is prominent before irradiation. (B) The spindle has disappeared shortly after localized, ultraviolet irradiation of cytoplasm. Courtesy R. B. Uretz.



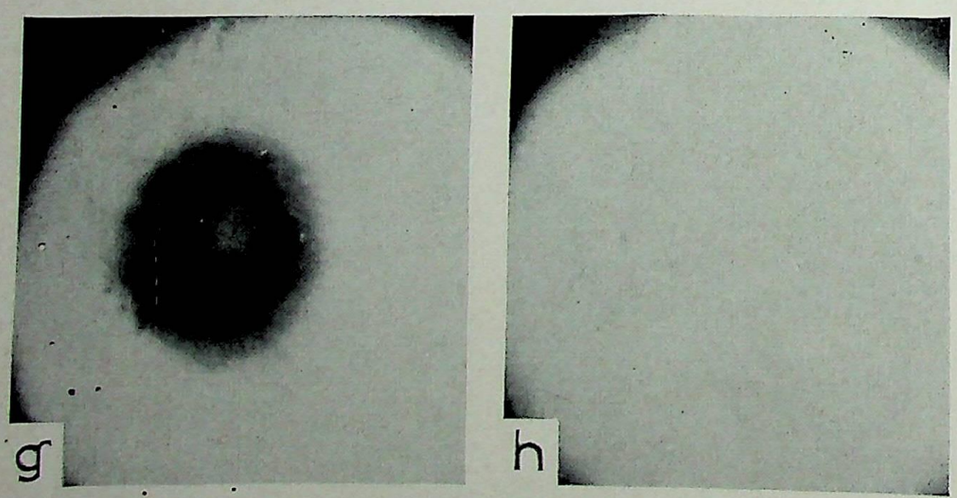


FIG. 11. Newt cell in prophase. (e) Before irradiation; (f) shortly after ultraviolet irradiation of 4-μ diam circle of nucleus. The paled area at the place of irradiation is shown by the arrow. (g) After irradiation, photographed by 260 mμ light, and (h) photographed by 310 mμ light. The diminished absorption of the paled area at 260 mμ is obvious (after R. P. Perry)[8].

general absorption spectrum typically shown by proteins containing aromatic amino acids. There is a minimum at 250 mμ and a broad maximum between 260 and 280 mμ.

The relative effectiveness of the two types of irradiation is roughly as follows: Irradiation with ultraviolet light of an 8-μ spot of a chromosome will require 5 sec to make it sticky and about 10 sec to make it pale, while spindle destruction will need about 150 sec of irradiation of an 8-μ spot of cytoplasm. Chromosomes irradiated locally with 2 Mev protons become sticky with 20 protons and pale with tens of thousands. It seems to require a million or more protons delivered to a small area of cytoplasm to destroy the spindle; we have not done many such experiments and this is a very approximate figure.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

SUMMARY

(1) Irradiation of dividing newt cells in culture with 200-kv x-rays or with ultraviolet light produces few or many abnormalities depending on the amount of radiation applied.

(2) Localized irradiation of chromosomes with a few tens of protons makes them sticky. But no abnormalities in the mitotic process result when a small area of cytoplasm is irradiated with 28 300 protons.

(3) Localized irradiation of cytoplasm with ultraviolet light prevents formation of the spindle or destroys it if it has already formed. In such cells, groups of whole chromosomes (not chromatids) move apart, usually in inequal numbers, to form the daughter nuclei of the dividing cell.

(4) The spindle thus seems to provide a mechanism for guiding the chromatids in anaphase, since, in the absence of a spindle, groups of whole chromosomes can move apart and form the nuclei for the daughter cells.

(5) Localized ultraviolet irradiation of chromosomes changes the index of refraction at the affected area. Such pale spots stain very faintly for DNA with the Feulgen method and absorb very little at 260 mμ.

BIBLIOGRAPHY

[1] R. E. Zirkle and W. Bloom, Science, **117**, 487 (1953).

[2] R. B. Uretz, W. Bloom, and R. E. Zirkle, Science, **120**, 197 (1954).

[3] R. B. Uretz and R. P. Perry, Rev. Sci. Instr. **28**, 861 (1957).

[4] W. Bloom, R. E. Zirkle, and R. B. Uretz, Ann. N. Y. Acad. Sci. **59**, 503 (1955).

[5] R. B. Uretz and R. E. Zirkle, Biol. Bull. **109**, 370 (1955).

[6] R. E. Zirkle, Advances in Biol. and Med. Phys. **5**, 103 (1957).

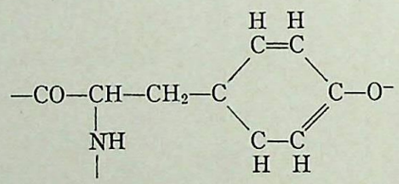
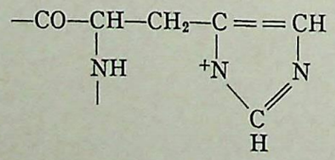
[7] S. Tschachotin, *Handbuch der biol. Arbeitsmethoden* (Urban and Schwarzenberg, Berlin, 1938), Division 5, Part 10, p. 877.

[8] R. P. Perry, Exptl. Cell. Research **12**, 546 (1957).

[9] W. Bloom and R. J. Leider, Radiation Research **3**, 214 (1955).

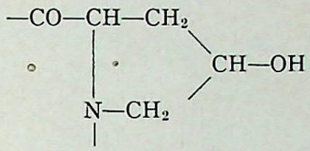
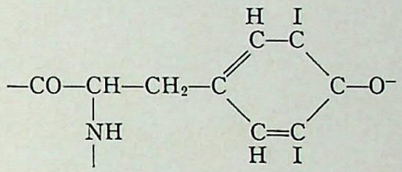
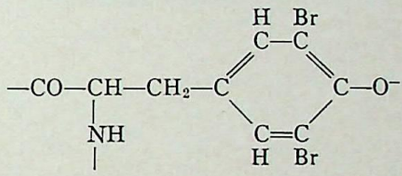
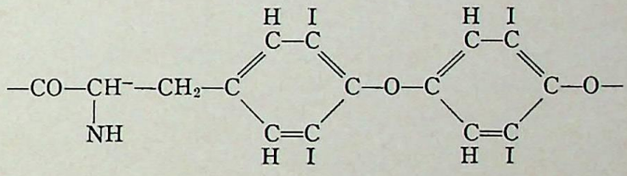
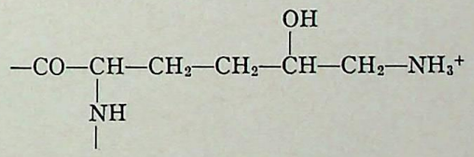
CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

TABLE I.—*Continued.*

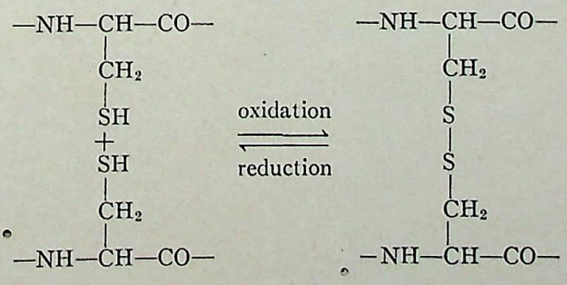
Residue symbol	(a) Residue name (b) Amino-acid name (trivial) (c) Amino-acid name (systematic)	Residue formula	
-cyS-* | -cyS-	(a) cystinyl (b) cystine (c) β-β′-dithiobis (α-aminopropionic acid)	—CO—CH—CH$_2$—S—S—CH$_2$—CH—NH— | NH | CO |	
NH$_2$ | -asp-	(a) asparaginyl (b) asparagine (c) α-aminosuccinamic acid	—CO—CH—CH$_2$—CO—NH$_2$ | NH |	
NH$_2$ | -glu-	(a) glutaminyl (b) glutamine (c) α-aminoglutaramic acid	—CO—CH—CH$_2$—CH$_2$—CO—NH$_2$ | NH |	
	Amino-acid residues usually found in proteins Acidic (ionized form)		
-OH-*	terminal carboxyl	—O$^-$ (—CO$_2^-$)	pK_a=3.6–4.0
-asp-	(a) aspartyl (b) aspartic acid (c) aminosuccinic acid	—CO—CH—CH$_2$—CO$_2^-$ | NH |	pK_a=3.9–4.7
-glu-	(a) glutamyl (b) glutamic acid (c) α-aminoglutaric acid	—CO—CH—CH$_2$—CH$_2$—CO$_2^-$ | NH |	pK_a=3.9–4.7
-tyr-	(a) tyrosyl (b) tyrosine (c) α-amino-β-(*p*-hydroxyphenyl) propionic acid		pK_a=8.5–10.9
-cySH-*	(a) cysteinyl (b) cysteine (c) α-amino-β-mercaptopropionic acid	—CO—CH—CH$_2$—S$^-$ | NH |	pK_a=*ca* 10
	Basic (ionized form)		
-his-	(a) histidyl (b) histidine (c) α-amino-β-imidazolyl-propionic acid		pK_a=6.4–7.0
-H-*	terminal amino	—H$_2^+$ (—NH$_3^+$)	pK_a=7.4–8.5
-lys-	(a) lysyl (b) lysine (c) α-ε-diaminocaproic acid	—CO—CH—CH$_2$—CH$_2$—CH$_2$—CH$_2$—NH$_3^+$ | NH |	pK_a=8.5–10.9
-arg-	(a) arginyl (b) arginine (c) α-amino-γ-guanidino-valeric acid	—CO—CH—CH$_2$—CH$_2$—NH—CH=NH$_2^+$ | NH | NH$_2$	pK_a=11.9–13.3

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

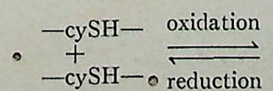
TABLE I.—*Continued.*

Residue symbol	(a) Residue name (b) Amino-acid name (trivial) (c) Amino-acid name (systematic)	Residue formula	
	Amino-acid residues occasionally found in proteins		
	Neutral		
-hypro-	(a) hydroxyprolyl (b) hydroxyproline (c) 4-hydroxypyrroli-dine-2-carboxylic acid (Found only in collagen)		
	Acidic (ionized form)		
	(a) diiodotyrosyl (b) diiodotyrosine (c) 3,5-diiodotyrosine (Found only in marine organisms and thyroglobulin)		$pK_a = 6.5$
	(a) dibromotyrosyl (b) dibromotyrosine (c) 3,5-dibromotyrosine (Found only in marine organisms)		$pK_a = ca\ 7$
	(a) thyroxyl (b) thyroxine (c) 3,3′,5,5′-tetra-iodothyronine (Found only in thyroglobulin)		$pK_a = ca\ 6.5$
	Basic (ionized form)		
-hylys-	(a) hydroxylysyl (b) hydroxylysine (c) α, -diamino-hydroxy-caproic-acid (Found only in collagen)		$pK_a = ca\ 11$

oxidized to form the disulfide linkage,



or



is found in all but a few proteins. This linkage may be between two otherwise separate peptide chains, or it may be an additional intrachain link. Both of these types of disulfide bonds are found in the insulin molecule, shown diagrammatically in Fig. 3. Calvin[4] has recently discussed the geometry of such C—S—S—C linkages. The actual spatial arrangement of a disulfide bond is shown in Fig. 4. The intrachain disulfide of insulin leads to a "link" in the polypeptide chain. This link contains six amino-acid residues (20 atoms), and it is of great interest to find that links of exactly the same number of residues are found in the peptide hormones oxytocin, arginine vasopressin, and lysine vasopressin (see Stetten, p. 563).

Recent studies by Perlmann[5] have done much to elucidate the various types of crosslinkages which involve phosphoric-acid residues. Orthophosphate —O—PO_2^-—O— and pyrophosphate —O—PO_2^-—

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

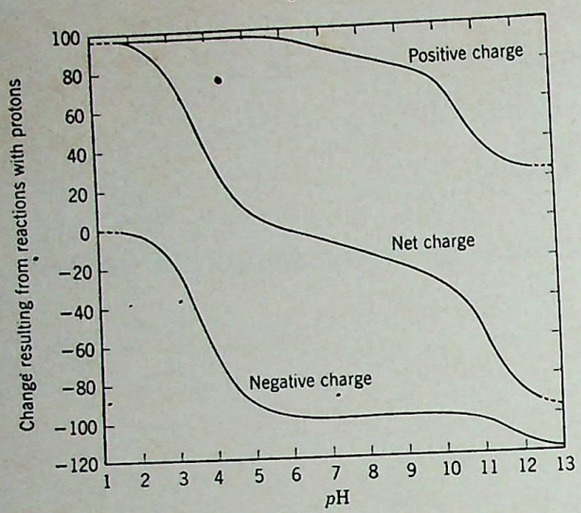
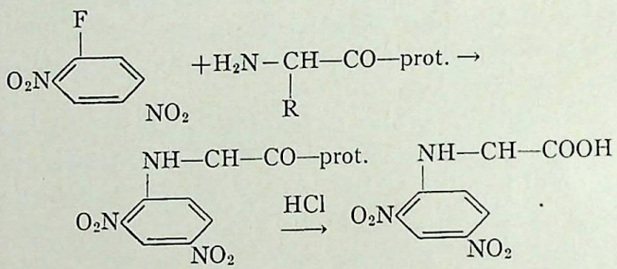


FIG. 5. Titration curve of bovine serum albumin. Also represented is the total positive charge and the total negative charge. These values are computed for 1 C-terminal carboxyl residue, 99 aspartyl and glutamyl residues, 19 tyrosyl residues, 1 cysteinyl residue, 16 histidyl residues, 1 N-terminal amino residue, 57 lysyl residues, and 22 arginyl residues.

figuration of these molecules. Each amino acid contains α-NH_3^+ and α-COO^- groups, and the acidic or basic amino acids also contain an additional charged group. When amino acids are combined in peptides or proteins, the α-NH_3^+ and α-COO^- groups are lost through the formation of the peptide bond, and only the terminal NH_3^+ and -COO^- residues can become ionized. But the acidic and basic groups of the amino-acid residues of category two (Table I) remain capable of being ionized, and it is primarily these residues which give the dipolar and ionic properties to the protein molecule. Titration of a protein solution with strong acid or base allows a calculation of the number of available acidic and basic groups and an estimation of their *p*K's. Such a titration curve for bovine serum albumin calculated from the measurements of Tanford[9] is shown in Fig. 5. It can be seen that, when the net charge of the protein is zero, there are approximately 94 positively charged groups (basic groups of histidyl, terminal amino, lysyl, and arginyl residues) and an equal number of negatively charged groups (carboxyl groups of terminal carboxyl, aspartyl, and glutamyl residues). The properties of a protein are greatly influenced by the presence of these many charged groups, and the dipole moment of the uncharged protein is a measure of the symmetry of the distribution of these charges. Studies of the titration curves of many proteins have indicated that most of the acidic and basic amino-acid residues are available for reaction. [Rice (p. 69) discusses some of the difficulties in the analysis of titration curves of proteins and other polyelectrolytes.] In certain instances, however, some of these residues appear to be unavailable for reaction unless the protein is taken to extremes of *p*H.[8] Two well-established examples of this behavior are seen in hemoglobin and in ribonuclease. In hemoglobin, about 36 groups (probably 18 carboxyl groups of aspartyl and glutamyl residues, and 18 ϵ-amino groups of lysyl residues) are found to be unreactive at neutral *p*H values,[10] whereas in ribonuclease, 3 of the 6 tyrosyl residues are found to be unreactive at their normal *p*K.[11]

One of the most exciting developments concerning the structure of proteins has been the evolution of methods which have led to the identification of the sequence of amino-acid residues in the peptide chains of a number of proteins. These developments have been made possibly largely through the work of Sanger and his colleagues who blazed the way by determining the sequence in insulin, a protein of about 6000 molecular weight.[12] Sanger's approach was to develop a method for the identification and estimation of the N-terminal residues of proteins and peptides. The reaction of the N-terminal residue with 1.2,4-fluorodinitrobenzene (FDNB) leads to the yellow dinitrophenyl (DNP) compound,



This reaction is carried out under mild (slightly alkaline) conditions where the peptide bonds are quite stable. Acid hydrolysis of the resulting DNP compound leads to a mixture of the various amino acids present in the protein and the DNP-amino acid from the N-terminal residue. The DNP-amino acids are reasonably stable under the conditions of acid hydrolysis, and can usually be obtained in good yield. There are certain DNP-amino acids, for example DNP-proline and DNP-cysteine, which are considerably less stable than others, however, and substantial corrections must be made for the destruction of such DNP-amino acids during hydrolysis.

If the DNP-protein is only partially hydrolyzed, DNP-peptides can be isolated, and subsequent complete hydrolysis of these purified DNP-peptides reveals the nature of the amino-acid residues and the N-terminal residue; whereas a partial hydrolysis of the purified DNP-peptide can lead to the arrangement of the residues in the N-terminal peptide. By this method, the four or five residues adjoining the N-terminal residue can be arranged in sequence. In the studies on insulin, the N-terminal sequences DNP-phe-val-asp-glu- and DNP-gly-ileu-val-glu-glu- were identified in this way. Also, this same method of attack has been used to determine the residue sequence in various purified peptides obtained from either acid or enzymatic hydrolysis of the protein. After a large number of such peptides were completely identified (about 65 in the case of the *B*-chain of insulin), it was found that there

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

TABLE II. Peptides identified in hydrolyzates of fraction B of oxidized insulin.[a]

Dipeptides from acid and alkaline hydrolyzates

H-phe-val-OH H-his-leu-OH H-his-leu-OH H-ala-leu-OH H-gly-glu-OH H-thr-pro-OH
H-val-asp-OH H-leu-cySO$_3$H-OH H-leu-val-OH H-leu-val-OH H-glu-arg-OH H-lys-ala-OH
H-asp-glu-OH H-cySO$_3$H-gly-OH H-val-glu-OH H-val-cySO$_3$H-OH H-arg-gly-OH
H-glu-his-OH H-ser-his-OH H-glu-ala-OH H-cySO$_3$H-gly-OH H-gly-phe-OH

Tripeptides from acid and alkaline hydrolyzates

H-phe-val-asp-OH H-leu-cySO$_3$H-gly-OH H-ala-leu-tyr-OH H-gly-glu-arg-OH H-pro-lys-ala-OH
H-val-asp-glu-OH H-ser-his-leu-OH H-tyr-leu-val-OH
H-glu-his-leu-OH H-leu-val-glu-OH H-leu-val-cySO$_3$H-OH
H-his-leu-cySO$_3$H-OH H-val-glu-ala-OH H-val-cySO$_3$H-gly-OH

Higher peptides from acid and alkaline hydrolyzates

H-phe-val-asp-glu-OH H-ser-his-leu-val-glu-OH H-tyr-leu-val-cySO$_3$H-OH H-thr-pro-lys-ala-OH
H-phe-val-asp-glu-his-OH H-ser-his-leu-val-glu-ala-OH H-leu-val-cySO$_3$H-gly-OH
H-glu-his-leu-cySO$_3$H-OH H-his-leu-val-glu-OH
H-his-leu-cySO$_3$H-gly-OH H-leu-val-glu-ala-OH
H-ser-his-leu-val-OH

Sequences deduced from above peptides

H-phe-val-asp-glu-his-leu-cySO$_3$H-gly- -tyr-leu-val-cySO$_3$H-gly- -thr-pro-lys-ala
-ser-his-leu-val-glu-ala- -gly-glu-arg-gly-

Peptides identified in peptic hydrolyzate

H-phe-val-asp-glu-his-leu-cySO$_3$H-gly-ser-his-leu-OH H-leu-val-cySO$_3$H-gly-glu-arg-gly-phe-OH
H-val-glu-ala-leu-OH H-tyr-thr-pro-lys-ala-OH
H-his-leu-cySO$_3$H-gly-ser-his-leu-OH

Peptides identified in chymotryptic hydrolyzate

H-phe-val-asp-glu-his-leu-cySO$_3$H-gly-ser-his-leu-val-glu-ala-leu-tyr-OH H-tyr-thr-pro-lys-ala-OH
H-leu-val-cySO$_3$H-gly-glu-arg-gly-phe-phe-OH

Peptides identified in tryptic hydrolyzate

H-gly-phe-phe-tyr-thr-pro-lys-ala-OH

Structure of the B-(phenylalanyl terminal) chain of insulin

H-phe-val-asp-glu-his-leu- (cyS-) -gly-ser-his-leu-val-glu-ala-leu-tyr-leu-val- (cyS-) -gly-glu-arg-gly-phe-phe-tyr-thr-pro-lys-ala-OH

[a] Taken from F. Sanger, Advances in Protein Chem. 7, 56 (1952).

was only one sequence which would fit all of the experimental results, assuming that there was a single polypeptide chain of about 30 residues. It was further assumed that no rearrangements of the residues had occurred during the hydrolyses.

The attack outlined in the foregoing is applicable only to single peptide chains devoid of interchain linkages, such as the disulfide link described earlier. Since insulin and most other proteins contain interchain and/or intrachain disulfide linkages, these must be broken before the sequence studies can be undertaken by the foregoing methods. In the case of insulin, oxidation of all of the disulfide linkages to sulfonic-acid groups was carried out with performic acid. This reagent thus converts cystine to cysteic acid (H—cySO$_3$H—OH), a strong acid. It also converts methionine to the corresponding sulfone, and tryptophan to unidentified products. Since insulin contains no methionine or tryptophan, performate oxidation was capable of converting the insulin to two chains, called the *A*-(glycine terminal) chain and the *B*-(phenylalanine terminal) chain, with each half-cystine residue (-cyS-) converted to a cysteic-acid residue (-cySO$_3$H-). The *A*- and *B*-chains were then purified, and the sequences of amino-acid residues in each of the two peptide chains were determined by the method outlined in the following. Table II records a number of the peptides obtained from such a study of the *B*-chain, and Fig. 6 shows the entire sequence for the insulin molecule. The acid hydrolysis used in the determination of the amino-acid sequence in insulin caused the liberation of six moles of ammonia, originally present as the amide groups of the

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

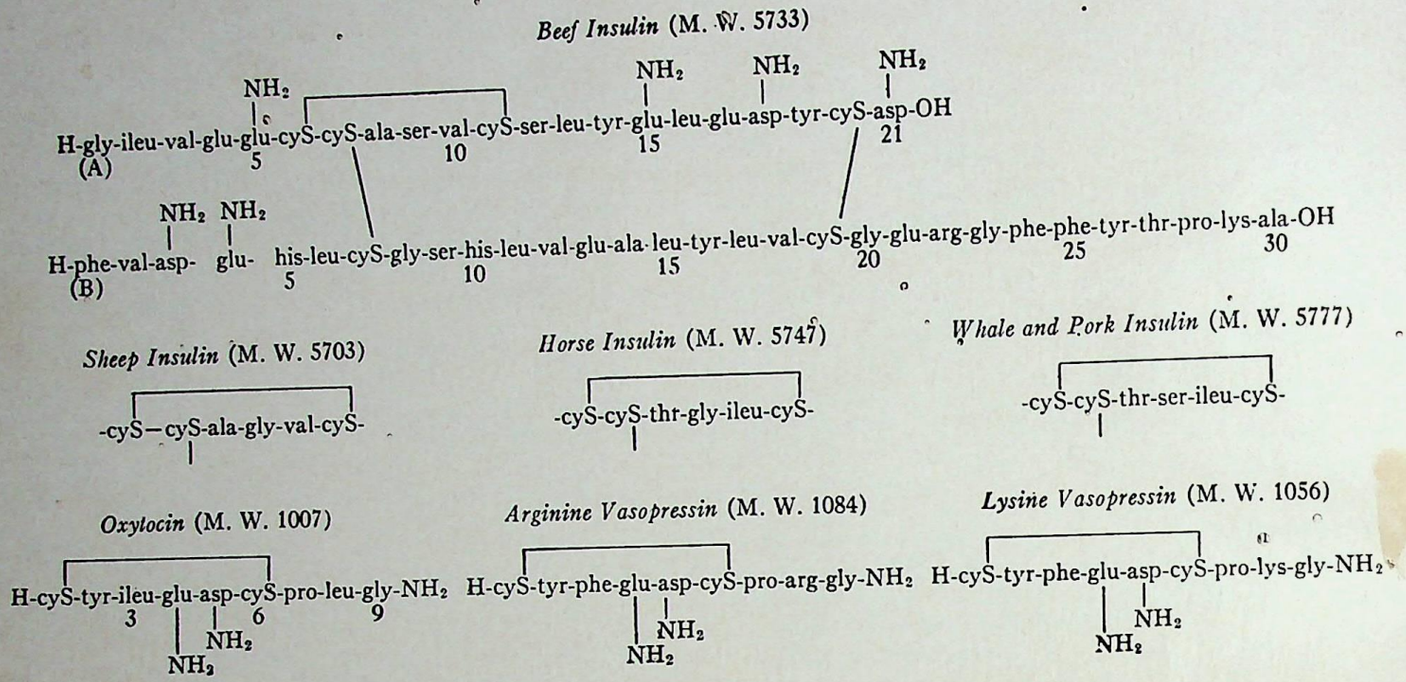


FIG. 6. Amino-acid sequences in insulin and some hormone peptides.

asparaginyl and glutaminyl residues. Reduction of the insulin by lithium borohydride ($LiBH_4$) in tetrahydrofuran converted the ω-carboxyl groups to primary alcohol groups, and acid hydrolysis of the reduced insulin then showed that three asparaginyl residues, three glutaminyl residues, and four glutamyl residues were present in insulin. In order to locate these amide groups on the individual glutamic-acid residues, two methods were used which depend on the fact that enzymatic hydrolysis does not break the ω-amide linkage. The purified peptides obtained after enzymatic hydrolysis could then be tested for amide content, either by studies of their electrophoretic behavior (since the amide derivatives had one negative charge less than the corresponding ω-acid derivatives), or by studies of the extent of ammonia liberation during their acid hydrolysis. In this way, it was found that residues 5, 15, 18, and 21 of the *A*-(glycyl terminal) chain and residues 3 and 4 of the *B*-chain were present in the amide form.

In order to find the distribution of the disulfide bridges, Sanger submitted intact insulin to partial enzymatic hydrolysis. The resulting cystine-containing peptides were then purified and sequences of these peptides were determined. This gave Sanger the final data required to completely identify the sequences and cross-linkages in beef insulin,[13] and resulted in the formula given in Fig. 6. The intrachain disulfide linkage found in the *A*-(glycyl terminal) chain is of special interest. The half-cystinyl residue at position 6 is combined with the position 11 half-cystinyl residue through a disulfide linkage. Studies on horse, sheep, whale, and pork insulin indicate sequences similar to those found in beef insulin, except for the amino acids in positions 8, 9, and 10 of the *A*-(glycyl terminal) chain. These residue differences are all within the hexapeptide disulfide link. Figure 6 also shows the sequences of the insulins of the other species studied. Whale and pork insulin are seen to be identical.

As mentioned in the discussion of the intrachain disulfide linkage, three peptide hormones isolated from the anterior pituitary and synthesized by du Vigneaud —oxytocin, arginine vasopressin, and lysine vasopressin —all contain a similar hexapeptide disulfide link. The formulas of these peptides are shown in Fig. 6 also, and the possible significance of this structure in terms of hormone structure is discussed by Stetten (p. 563).

The methods developed by Sanger and his colleagues for the study of the amino-acid sequence in insulin have subsequently been applied by a number of laboratories to the elucidation of the structure of several other proteins. Other methods of attack involving replacement of the DNP compounds by other types of derivatives have been developed, and the stepwise freeing of N-terminal and C-terminal residues by enzymatic hydrolysis catalyzed by aminopeptidases and carboxypeptidase has proved to be fruitful. Methods have also been developed to replace the performate oxidation procedure for the elimination of disulfide bonds, and these can be applied in the study of proteins containing tryptophan and other easily oxidized amino acids. A short summary of these methods for amino-acid sequence determinations can be found in a recent review by Anfinsen and Redfield.[14]

Figures 7–9 summarize some of the more complete amino-acid sequence studies on other proteins and large peptides. The most complete amino-acid sequence study of a series of related hormones is the work on the melanocyte-stimulating hormones (MSH) and adrenocorticotropic hormones (ACTH). Figure 7 records the

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

Beef Melanotropin (*Seryl-β-MSH*) (M.W. 2134)

H-asp-ser-gly-pro-*tyr*-lys-*met-glu-his-phe-arg-try-gly*-ser-*pro*-pro-lys-asp-OH
5 10 15 18

Pork β-Melanocyte-Stimulating Hormone (*Glutamyl-β-MSH*) (M.W. 2176)

H-asp-glu-gly-pro-*tyr*-lys-*met-glu-his-phe-arg-try-gly*-ser-*pro*-pro-lys-asp-OH
5 10 15 18

Pork α-Melanocyte-Stimulating Hormone (*α-MSH*) (M.W. 1665)

CH_3-CO-ser-*tyr*-ser-*met-glu-his-phe-arg-try-gly*-lys-*pro*-val-NH_2
5 10 13

Sheep α-Corticotropin (*α-ACTH*) (M.W. 4540)*

H-ser-*tyr*-ser-*met-glu-his-phe-arg-try-gly*-lys-*pro*-val-gly-lys-lys-arg-arg-pro-val-
5 10 15 20

NH_2
|
-lys-val-tyr-pro-ala-gly-glu-asp-asp-glu-ala-ser-glu-ala-phe-pro-leu-glu-phe-OH
21 25 30 35 39

ACTH Preparation	Residue No. 25 26 27 28 29 30 31 32 33
Sheep α-corticotropin	-ala-gly-glu-asp-asp-glu-ala-ser-glu-(NH_2)
Pork corticotropin-A (Cyanamid)	-asp-gly-ala-glu-asp-glu-leu-ala-glu-
Pork corticotropin-A (Armour)	-gly-ala-glu-asp-asp-glu-leu-ala-glu-

FIG. 7. Amino-acid sequences in melanocyte-stimulating and adrenocorticotropic hormones. Sheep α-corticotropin contains one more amide group, probably in position 27, 28, or 29. Beef α-corticotropin appears from preliminary structural investigations to be identical with sheep α-ACTH. All of these ACTH preparations seem to have no loss of biological potency when residues 29 to 39 (11 residues) are removed by limited enzyme hydrolysis.

results of extensive studies on these hormones in some five or six laboratories. Recent reviews of this work have been presented by Li,[15–17] Harris,[18] and by Anfinsen and Redfield.[14] The amino-acid sequences of all of the MSH and ACTH materials studied to date have shown almost identical sequences . . .-*tyr*-.X.-*met-glu-his-phe-arg-tyr-gly*-.Y.-*pro*-. . . . This sequence has been shown in italics in Fig. 7, and is seen to occur at different distances from the N-terminal residue in the β-MSH preparations (residues 5–15) and in α-MSH and the various ACTH preparations (residues 2–12). The residues X and Y in this unique sequence are seen to be lysyl or seryl, and seryl or lysyl, respectively (see Stetten, p. 563). The two pork corticotropins-A, as prepared in the Cyanamid and in the Armour Laboratories, are known to differ with respect to the total content of amide groups (none for the Armour product, and one for the Cyanamid product). The sequence differences in residues 25–28 of the two corticotropin-A preparations have not been definitely shown to indicate different sequences in the two preparations, since technical difficulties may have occurred in the sequence determination. The sheep α-corticotropin preparation of Li differs from the corticotropin-A, in that Li's preparation contains two amide groups (one is glutaminyl residue 33, and one not yet located, but probably in position 27, 28, or 29), and also one more seryl and one less leucyl residue. It may be noted that the component amino acids in positions 25–28 are the same in all three ACTH preparations, but differ in order. The different location of the amide residue (position 33 in sheep α-corticotropin and 30 in pork corticotropin-A) appears to represent a real difference in sequence. Preliminary structural studies by Li have indicated that the amino-acid sequence for beef α-corticotropin is identical with that for sheep α-corticotropin.[17]

The two β-MSH preparations are seen to differ only in the second residue, seryl for the beef hormone and glutamyl for the pork hormone. These materials are, therefore, often referred to as seryl-β-MSH and glutamyl-β-MSH. The acetylated amino group of the N-terminal seryl residue in α-MSH has only recently been identified,[18] and its presence is rather unusual. The presence of this group, as well as of the amide group of the C-terminal valyl residue in α-MSH, causes a drastic change in the hormonal activity of this compound as compared with that of the longer α-ACTH peptide.[18] This effect, as well as other aspects of the relation between the structure of these hormones, and their physiological activities, is discussed by Stetten (p. 563).

The amino-acid sequence in glucagon, a small protein involved in glucose metabolism, has recently been determined and is shown in Fig. 8. This protein, like the ACTH and MSH preparations, contains no cystinyl residues. Unlike these hormones, however, the sulfur-containing methionyl residue is present in glucagon. The relationship between glucagon and insulin in glucose metabolism is discussed by Stetten (p. 563). The

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

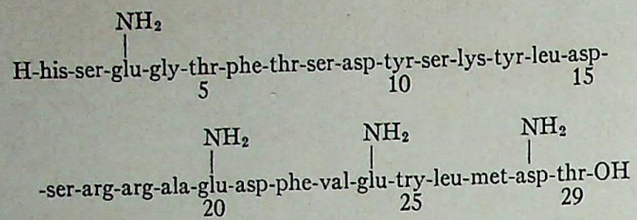


FIG. 8. Amino-acid sequence in glucagon (M. W. 3647).

sequence arrangements in all of the unrelated proteins so far evaluated show no more similarity than could be expected from simple probability considerations, and none of them indicates that repeating sequences are important elements in the structure of these small globular proteins. Whether or not repeating sequences occur in the larger globular proteins is as yet unknown. Some strong evidence of repeating elements has, however, been found in certain fibrous-protein structures, and in histone and protamine, proteins found associated with nucleic acid in the nucleoproteins.[14]

Ribonuclease is the most complicated protein in which an almost complete amino-acid sequence has been determined. This enzyme, containing 124 amino-acid residues and four disulfide bonds (molecular weight 12 000), contains more than twice the number of residues found in insulin. The presently known sequence as shown in Fig. 9 has recently been discussed by Hirs *et al.*[19] and by Anfinsen.[14,20] About 23 residues remain to be located definitely in the sequence. No unusual partial sequences seem to occur, with the possible exception of a number of repeats (-ala-ala-ala- in positions 4–7, and -met-met- in positions 29–30). These partial sequences can be compared with the unusual sequence -phe-phe-tyr- in the *B*-chain of insulin (positions 24–26).

Short, partial-residue sequences have been determined for a large number of proteins. A few of the more interesting partial sequences are reported in Figs. 10 and 11. The partial sequence shown for beef cytochrome-*c* (Fig. 10) is interesting in that this sequence contains the residues to which the porphyrin-*c* prosthetic group is attached. The exact attachment of the porphyrin-*c* residue may be either that shown in Fig. 10, or that of an otherwise identical structure where the porphyrin-*c* residue is rotated about the plane of the page by 180°. The corresponding partial sequences in a series of cytochrome-*c* preparations from sources other than beef are also shown in Fig. 10. Here, alanyl residues are sometimes replaced by seryl or glutamyl residues, lysyl residues by arginyl residues, valyl residues by lysyl residues, and glutaminyl residues by threonyl residues. The italicized sequence *-cyS-*.A.-.B.*-cyS-his-thr-val-glu-* has been found to occur in this order in each of the six cytochrome-*c* molecules. This study, involving work by Theorell, Tuppy, and others, has been reviewed by Tuppy[21] and by Anfinsen and Redfield.[14] The partial sequences shown for lysozyme and for human serum albumin (Fig. 11) are very incomplete. Many disulfide bonds are known to be present in these molecules, and serum albumin is somewhat unusual in that it contains a single cysteinyl residue. Many other partial sequences have been found for egg-white lysozyme,[14,22] but a study of these sequences shows that certain of them must be spurious, and it has not been possible to locate the various partial sequences in relation one to another. Anfinsen and Redfield[14] suggest that the study of lysozyme illustrates the limit of structural information that can be obtained by acid hydrolysis alone, and that more specific degradative methods must be applied before the complete amino-acid sequence can be obtained

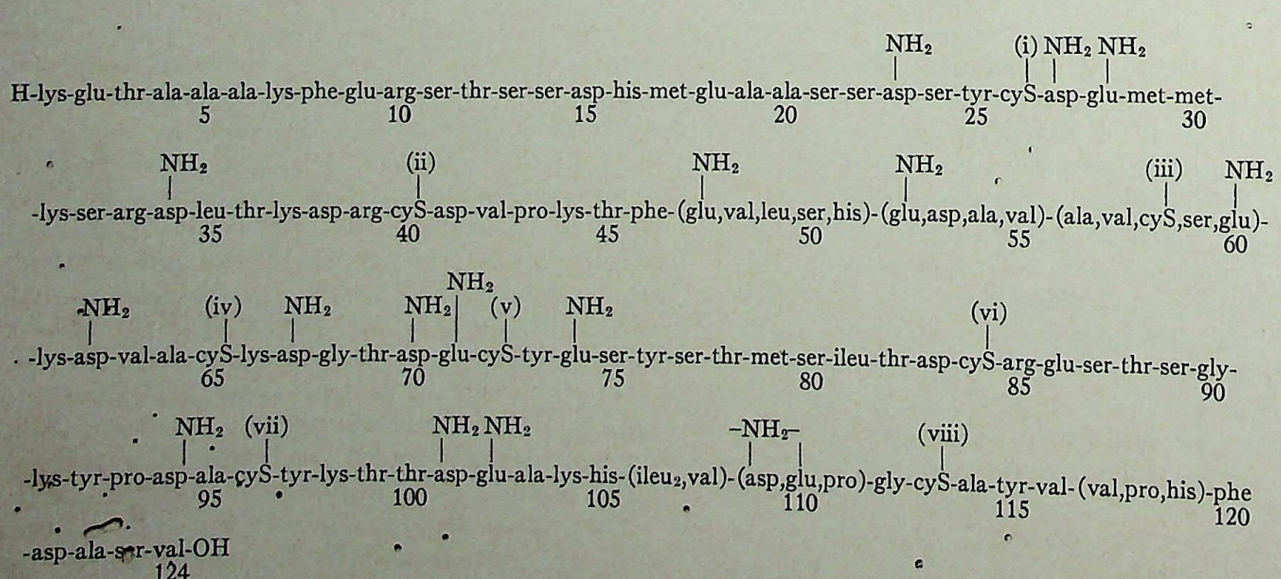
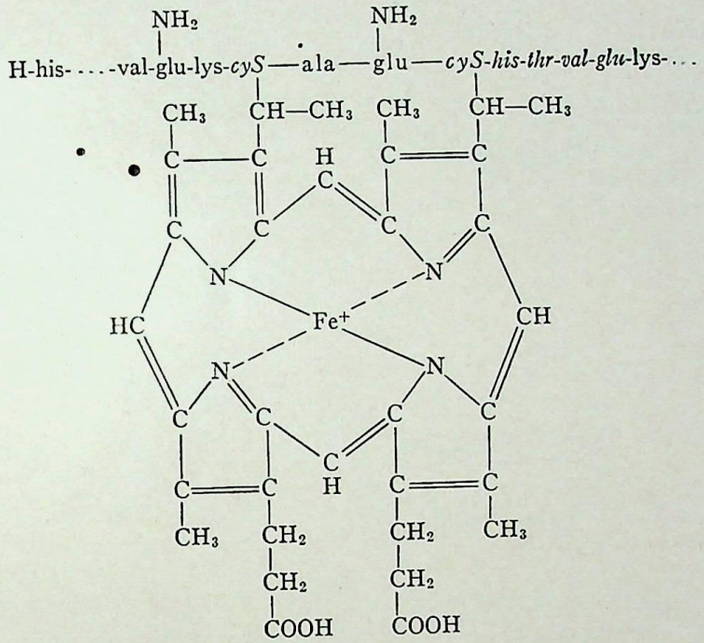


FIG. 9. Amino-acid sequence in ribonuclease (M. W. 12 000). Where the composition and position of a group of residues is known, but not the relative positions of the individual residues within the group, the listing of the group is enclosed in parentheses. The half-cystinyl residues are identified by roman numerals. Work of Anfinsen[20] has shown that the four disulfide bridges in ribonuclease are paired as follows: i–vi, iii–vii, ii–viii and iv–v (after studies by C. H. W. Hirs *et al.*[19]).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

Other species of serum albumin have been studied, and it has been shown that bovine serum albumin, while containing the same N-terminal residue (H-asp-), has threonyl as the second residue and ends with a different C-terminal sequence, possibly -(ala, leu, thr, val, ser)-ala-OH.[14] Porter[23] has obtained a fragment of the bovine serum-albumin molecule after mild chymotryptic hydrolysis which seems to represent about one-fifth of the entire molecule (molecular weight about 12 000).

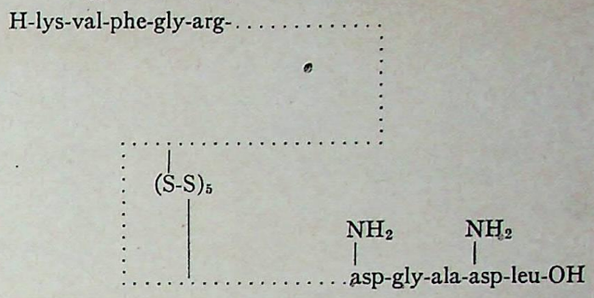
Beef Cytochrome-c (M.W. 13 100)



Preparation	*Observed residue sequence*
Beef	-val-glu-lys-*cyS*-ala-glu-*cyS-his-thr-val-glu*-lys-
Horse and pork	-lys-*cyS*-ala-glu-*cyS-his-thr-val-glu*-lys-
Salmon	-val-glu-lys-*cyS*-ala-glu-*cyS-his-thr-val-glu*-
Chicken	-val-glu-lys-*cyS*-ser-glu-*cyS-his-thr-val-glu*-
Silkworm	-val-glu-arg-*cyS*-ala-glu-*cyS-his-thr-val-glu*-
Yeast	-phe-lys-thr-arg-*cyS*-glu-leu-*cyS-his-thr-val-glu*-

FIG. 10. Partial amino-acid sequence in cytochrome-*c*. The exact attachment of the porphyrin-*c* residue may be either the residue shown or the structure where the porphyrin-*c* residue is rotated by 180° [from H. Tuppy in *Symposium on Protein Structure*, A. Neuberger, editor (Methuen and Company, Ltd., London; John Wiley and Sons, Inc., New York, 1958), p. 71].

Egg-white Lysozyme (M.W. 14 900)



Human Serum Albumin (M.W. 68 000)

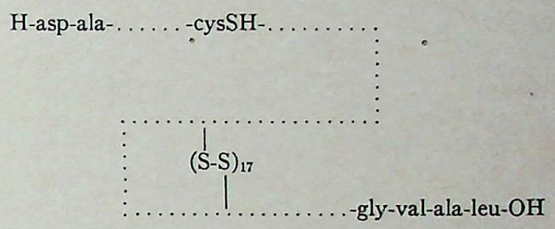


FIG. 11. Partial amino-acid sequences in lysozyme and serum albumin [from C. B. Anfinsen and R. R. Redfield, Advances in Protein Chem. **11**, 1 (1956)].

This fragment appeared to have very nearly the same configuration as part of the original albumin molecule, since it effectively inhibited the reaction of a specific antiserum to bovine serum albumin with its antigen (see Kauzmann, p. 549). This fragment contained a single N-terminal amino acid (H-phe-), the unique cysteinyl residue, and a single disulfide bond.

Further structural studies of amino-acid sequences are reported in other contributions to the Paris *Symposium on Protein Structure*.[24] Considerable sequence data exist for many other proteins, notably papain, growth hormone, trypsin, chymotrypsin, pepsin, hemoglobin, and tobacco mosaic virus.

It is necessary to always bear in mind the difficulties in the isolation of purified and homogeneous components in their native state. Proteins occur in nature as complex mixtures, and are often found in the presence of highly charged polysaccharides and/or other macromolecules. These naturally occurring colloidal mixtures are often enclosed by membranes of varying stability. The extraction of a particular protein from such a mixture is always difficult, and is often impossible without the use of procedures which may cause permanent changes in the configuration or even in the covalent linkages of the resulting protein preparation. A number of useful methods for the isolation and purification of such protein systems are currently available for the separation of these components. These methods include precipitation and differential extraction by high concentrations of salts or organic solvents, adsorption and partition between immiscible solvents, as well as such physical methods as electrophoresis and ultracentrifugation. A discussion of these methods cannot be undertaken here, but it is important to remember that the purified components isolated by these methods may

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

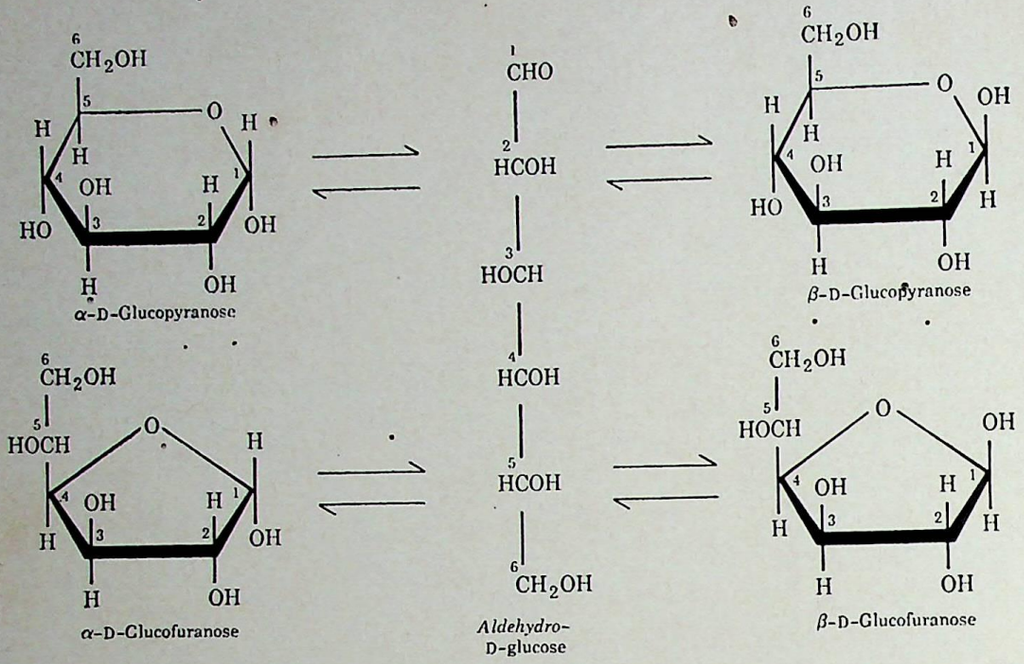
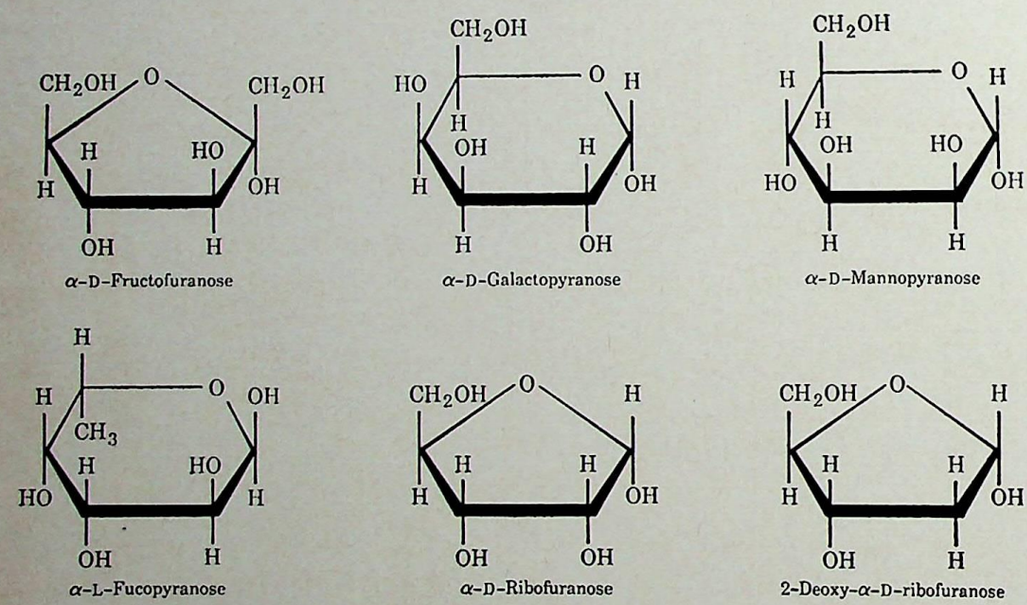


FIG. 12. Configuration of some monosaccharides.



not be truly homogeneous, and artifacts may be introduced by the isolation procedures.

Tests for the homogeneity of the purified components must be evaluated critically. Physical methods are most useful; they include studies of behavior in the ultracentrifuge and in electrophoresis, measurements of solubility (akin to the phase-rule methods of constant melting or boiling points as applied in the case of smaller molecules), and determination of the distribution coefficient between various solvents. Chemical methods showing the constancy of composition (in regard to the individual amino-acid residues, for example) are also of much value. Biological assay methods measuring activities ascribed to various components and quantitatively evaluating the activities of the purified components in terms of the activity of the native tissues are of special importance.

CARBOHYDRATES

Carbohydrates comprise another of the major groups of naturally occurring organic materials. They are often found in combination with proteins as glycoproteins and mucoproteins, especially in the higher animals. They also occur in combination with lipids to give cerebrosides, and in combination with purines and phosphoric acid to give nucleotides and nucleic acids. They form the structural elements of plants, bacteria, and certain animals in the forms of cellulose, chitin, and other high molecular-weight polysaccharides. They provide a means for the storage of chemical energy in the forms of amylose and amylopectin in plant starches, and glycogen in animals.

Although most monosaccharides are fairly stable when in the crystalline condition, they undergo many

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

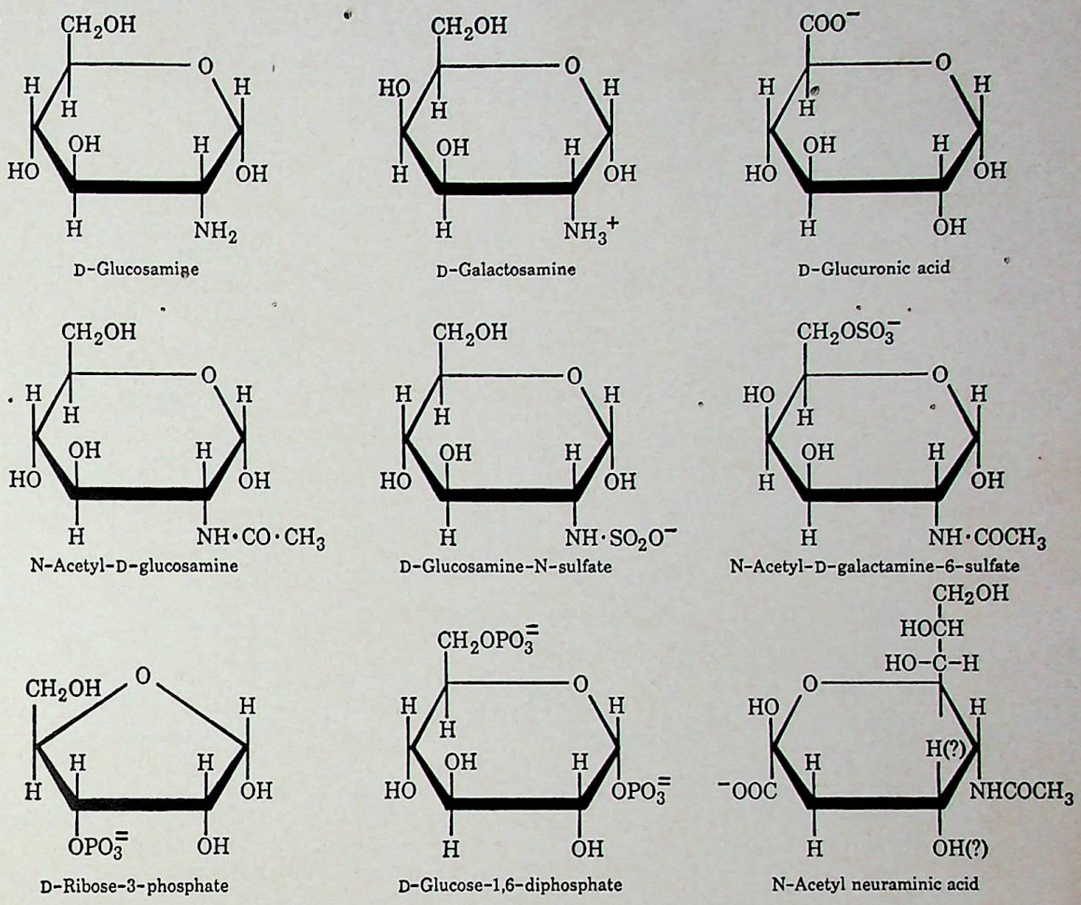


FIG. 13. Configuration of some charged sugar derivatives.

transformations when dissolved in water, particularly in the presence of acids or bases. Although the chemical formulas of the monosaccharides are often shown in a linear form, much evidence suggests that, in the main, they exist in the form of five- and six-membered rings called the furanose and pyranose forms. Each of these forms exists as α- and β-isomers, termed anomers, and interconversions between anomers and between ring isomers occur even under the mildest possible conditions of acidity and temperature. One of the simplest methods for demonstrating and studying the equilibria between the various forms is by the measurement of changes of optical rotation with time ("mutarotation"), which can easily be observed with freshly prepared sugar solutions. These various equilibria (for D-glucose) are illustrated in Fig. 12. The linear forms are shown in the Fischer formulation where the carbon atoms must be thought of as tetrahedrons, bonded together by angles projecting into the page and with the H— and —OH groups projecting out from the page. The ring forms shown in the Haworth formulation are to be viewed as a perspective representation with the heavy bonds extending out from the page. Side chains in the Haworth formulas are written according to the Fischer convention. The rings, shown here as planar, are oversimplified, since the valence angles in a single coplanar ring would be appreciably greater than those in a "strainless" structure having valence angles of 109°. The usual conformation of the pyranose ring is probably a chain form, but the free sugars in solution probably can occur to some extent in any of the possible conformations, since there is a spontaneous equilibrium with the aldehyde form. Derivatives in which the ring is fixed (because of substitution on C-1) probably are stabilized in the chair form.[25]

The aldehydo-hexoses such as D-glucose and D-galactose contain four asymmetric carbon atoms, while the keto-hexoses such as D-fructose and aldehydo-pentoses such as D-ribose contain three. Thus, many optical isomers exist (16 for the aldehydo-hexoses and 8 for the keto-hexoses and aldehydo-pentoses), and a number of these are found in various natural products. The ring structures contain an additional center of asymmetry at carbon atom 1, and the anomers of each ring structure are usually designated α- when the hydroxyl group of carbon atom 1 is *cis*- to the hydroxyl group at carbon atom 2, and β- when these two hydroxyl groups are *trans*-. In the case of 2-deoxy-D-ribose, the anomeric forms are designated like the D-ribose forms. α-D-mannose does not follow the (*cis*-) (*trans*-) rule given in the foregoing, and the commonly used nomenclature of Hudson is to be found in Pigman's review.[25] The designation L- indicates that the asymmetric carbon atom most remote from the reference group (e.g., aldehyde, keto, carboxyl, etc.) has the same configuration as L-glyceraldehyde (see Fig. 1). The L- and D-forms differ in the configuration of all asymmetric carbon atoms. In addition to the three monosaccharides

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

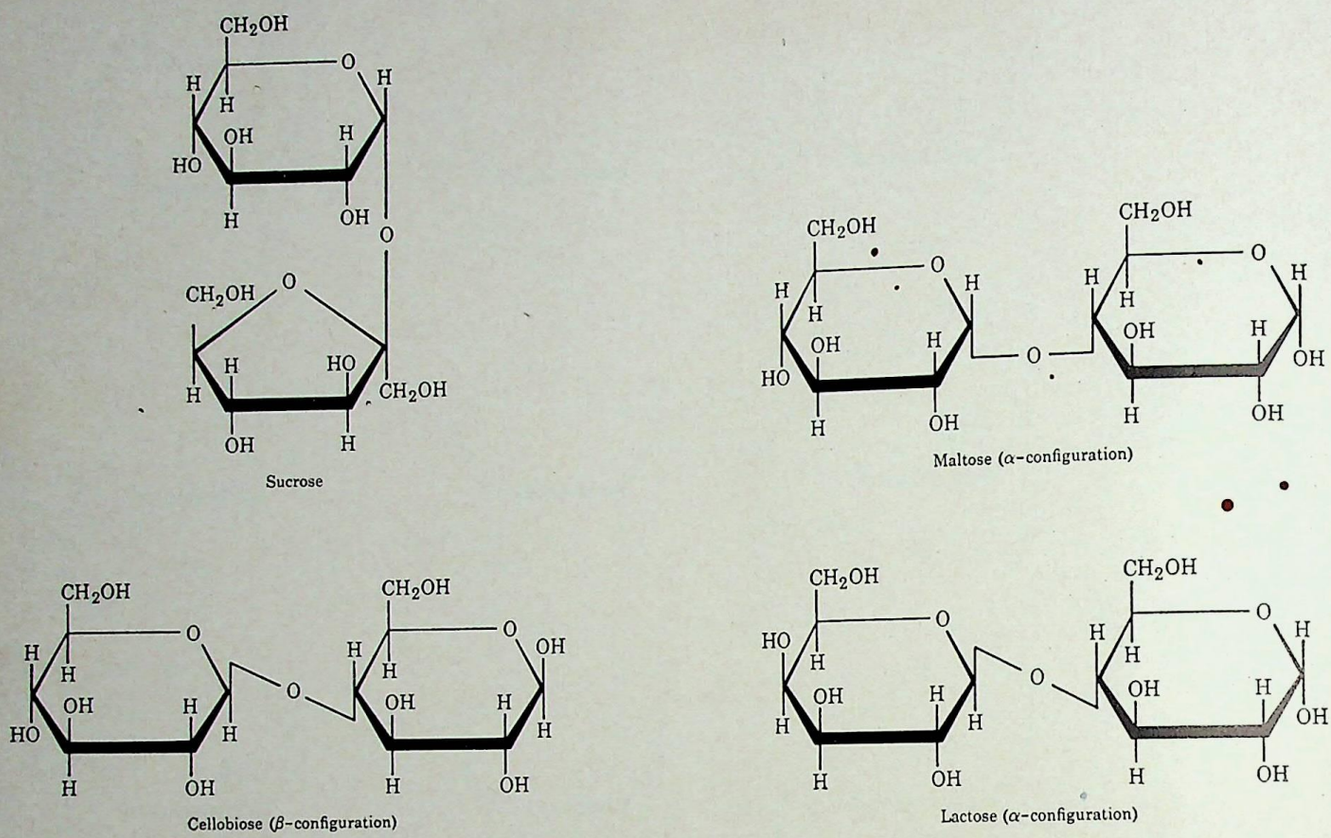


FIG. 14. Configuration of some disaccharides.

mentioned above, those most commonly found in bacterial and animal sources include D-galactose (like D-glucose with the configuration of C-4 reversed) and D-mannose (like D-glucose with configuration of C-2 reversed). Three other monosaccharides of major interest include L-fucose (6-deoxy-L-galactose), L-rhamnose (6-deoxy-L-mannose), and 2-deoxy-D-ribose, where the designation -deoxy- indicates that the appropriate —OH is replaced by —H.

Haworth formulas for most of the aforementioned monosaccharides are shown in Fig. 12. Each of these sugars would exist in the same sorts of conformations as are shown for D-glucose, and the ring structures commonly found in oligo- and poly-saccharides are portrayed in the diagrams. Only the α-anomer is depicted. The relative amounts of the four-ring conformations which will exist at equilibrium in aqueous solutions of the various monosaccharides vary considerably. In the case of neutral solutions of D-glucose, nearly two-thirds of the sugar seem to exist as β-glucopyranose, and just over one-third as α-glucopyranose. The two glucofuranose forms and the *aldehydo*-glucose form are thought to exist in minor amounts. In the case of D-mannose, over two-thirds exist as α-mannopyranose and just under one-third as β-mannopyranose. Solutions of D-ribose probably contain considerable quantities of the furanose and *aldehydo*- forms, and the mutarotation is complex, and exhibits a minimum.

There is also a number of biologically important derivatives of the monosaccharides, a few of which are listed in Fig. 13. Certain of these derivatives contain ionizable groups. Thus, glucosamine and galactosamine provide primary amine groups with a pK_a near 7.8. Glucuronic acid and galacturonic acid provide carboxyl groups with pK_a about 3.3; N-acetyl-neuraminic acid (sialic acid) has a stronger carboxyl group with pK_a 2.7; and the phosphate esters have two potential hydrogen ions, with pH_a values from 0.9 to 1.5 and from 5.9 to 6.3. The sulfate group of the amino-N-sulfate and the sulfate esters have a very strongly acidic hydrogen with $pK_a < 1$. The phosphate esters of the monosaccharides are products of intermediary metabolism, discussed in a number of other papers (see Lehninger, p. 136; Calvin, p. 147; Roberts, p. 170; Meister, p. 210).

The monosaccharide units of disaccharides may be alike as in maltose and cellobiose, or different as in sucrose and lactose. The hydrolysis of maltose and of cellobiose yields two molecules of D-glucose, whereas sucrose yields D-glucose and D-fructose, and lactose yields D-glucose and D-galactose. The monosaccharide residues are combined through an oxygen bridge of the hemiacetal hydroxyl to a second hydroxyl from another residue. The possible number of combinations of two monosaccharide units is large, since there are usually four or five free hydroxyl groups in the monosaccharide, and the oxygen bridge can come from either the α- or β-anomer of the other sugar unit. Figure 14 shows the

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

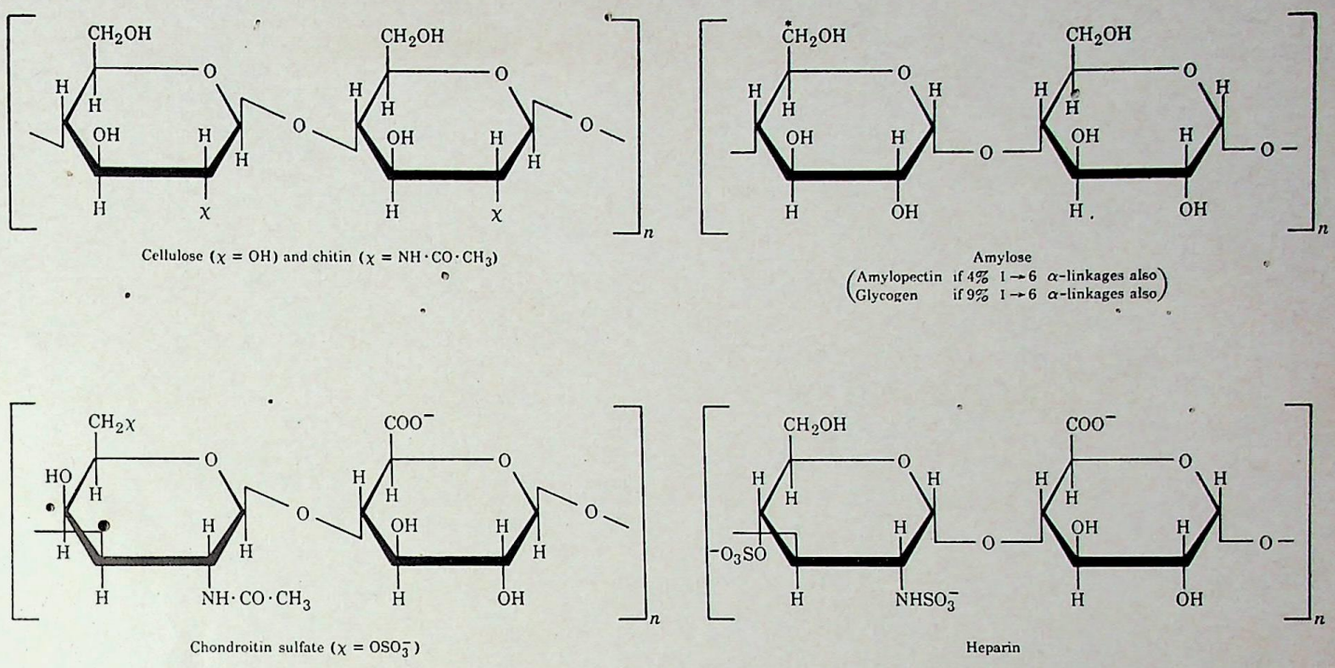


FIG. 15. Structural elements in selected polysaccharides.

structure of four common disaccharides, and illustrates some of the linkages possible. In sucrose, the two hemiacetyl hydroxyls are combined through an oxygen bridge between the two anomeric carbons, with the D-glucose being in the α-pyranose configuration, and the D-fructose in the β-furanose form. In lactose, the hemiacetyl hydroxyl (C-1) of β-D-galactopyranose bridges to the C-4 hydroxyl of D-glucopyranose. Both maltose and cellobiose contain 1 → 4 oxygen bridges between D-glucopyranose units, but in maltose there is an α-linkage while in cellobiose there is a β-linkage.

There are numerous polysaccharides of biological importance, with monosaccharide units or their derivatives as the polymer unit, often repeating either a single unit, ...A-A-A-A-..., or alternating two units, ...A-B-A-B-A-B-.... The type of glycosidic linkage between these units has the same diversity as was found in the disaccharides. There are certain of the polysaccharides that show a more complex structure owing to branching of the polymer chain by means of linkages through a third hydroxyl group of the polymer unit. Usually, the naturally occurring polysaccharides contain many hundreds of residues. The lability of the linkages makes the isolation and purification of undegraded materials difficult, and many polysaccharide preparations thus show molecular weights and polydispersity not characteristic of the native product.

Cellulose and amylose provide two of the simplest polysaccharide structures (Fig. 15). Both of these polymers are made up of D-glucose residues in the pyranose conformation, with 1 → 4 linkages between the units. In cellulose, the pyranose ring has the β-configuration and each linkage is like that in the disaccharide cellobiose; in amylose, the rings are in the α-configuration and each glycosidic linkage is like that in maltose. Chitin resembles cellulose, and is a polymer of N-acetyl-D-glucosamine units linked by a 1 → 4 β-glycosidic bond. This 1 → 4 β-glycosidic linkage leads to polymers of low solubility, whereas the 1 → 4 α-glycosidic linkage gives polymers of much higher solubility, owing to a spiraling of the macromolecule in a helix-like fashion induced by this type of bond.

Amylopectin and glycogen have a basic structure like amylose, but about five percent of the residues branch by means of 1 → 6 α-glycosidic bonds in the amylopectins, and about nine percent in the glycogens. The branched structure leads to solutions of much lower viscosity than would occur with amylose macromolecules of the same molecular weight. Detailed structures of glycogen and amylopectin have been established by the use of specific enzymes which differentiate between the 1 → 4 and 1 → 6 linkages, and Fig. 16 shows a representation of a segment of a glycogen molecule as indicated by such studies.

The aforementioned polysaccharides have been made up of uncharged polymer units. A more reactive class of polymers is made up from sugar derivatives. Less is known of their detailed structure, because the high charge density in the polymer makes the application of physicochemical methods much more difficult, and because the modifying groups are split from the polymerizing unit by many of the same reagents that are used to degrade the polysaccharide. Among the more important materials of this type, one finds hyaluronic acid, heparin, and chondroitin sulfate. These highly reactive polysaccharides are found in various animal tissues, and the structures now thought most likely are recorded in Fig. 15. None of these structures has been

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

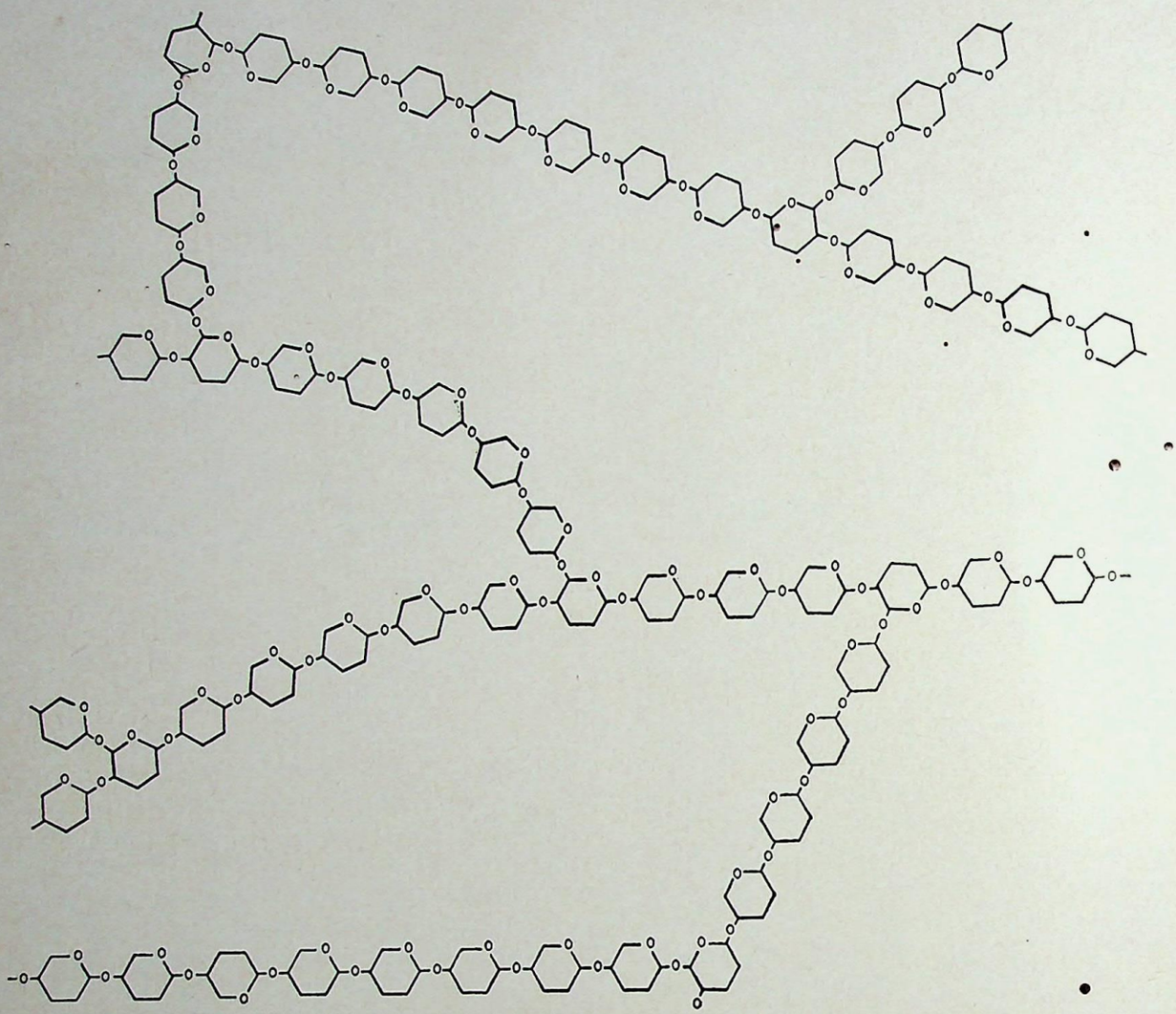


FIG. 16. Representation of a segment of a glycogen or amylopectin macromolecule.

definitely established at the present time. Other polysaccharides of this general type are to be found in bacterial systems, but none of these has structures that have been completely elucidated.

In the glycoproteins and mucoproteins, one finds large amounts of carbohydrate combined with protein.[26,27] Thus, the α_1-acid glycoprotein, a homogeneous crystallized material contains about 42% of carbohydrate with the following composition (expressed as moles per mole of glycoprotein, molecular weight 45 000): N-acetyl-D-glucosamine, 32; N-acetyl-nuraminic acid, 16; D-galactose, 18; D-mannose, 18; L-fucose, 2. Recent studies by Eylar indicate that most of this carbohydrate moiety can be removed in association with a small peptide fragment containing glutamyl, seryl, threonyl, and isoleucyl residues. Earlier studies by E. Smith showed that the smaller carbohydrate moiety of serum γ-globulin could be removed in association with a peptide fragment containing glutamyl, aspartyl, and seryl residues. These interesting results indicate that covalent linkages exist between the carbohydrate residues and certain of the aforementioned amino-acid residues.

LIPIDS

The classification "lipids" is taken to embrace the "fatty acids," all actual or potential esters of fatty acids, and often includes other materials soluble in "fat solvents," such as triterpenes, carotenoids, and fat-soluble vitamins. This large field is only touched upon here, since a recent short review by Lovern,[28] as well as the comprehensive reference book by Deuel[29] provide adequate background material. The simple fundamental lipid structures are often found to occur in complex structures of intermediate molecular weight, but to date no really high molecular-weight lipids have been discovered. On the other hand, these complex lipid molecules are often found in association with polysaccharides and/or proteins to form high molecular-weight materials.[30]

Although complex lipids often contain no charged groups (e.g., triglycerides, cholesterol, cholesterol

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by eGangotri

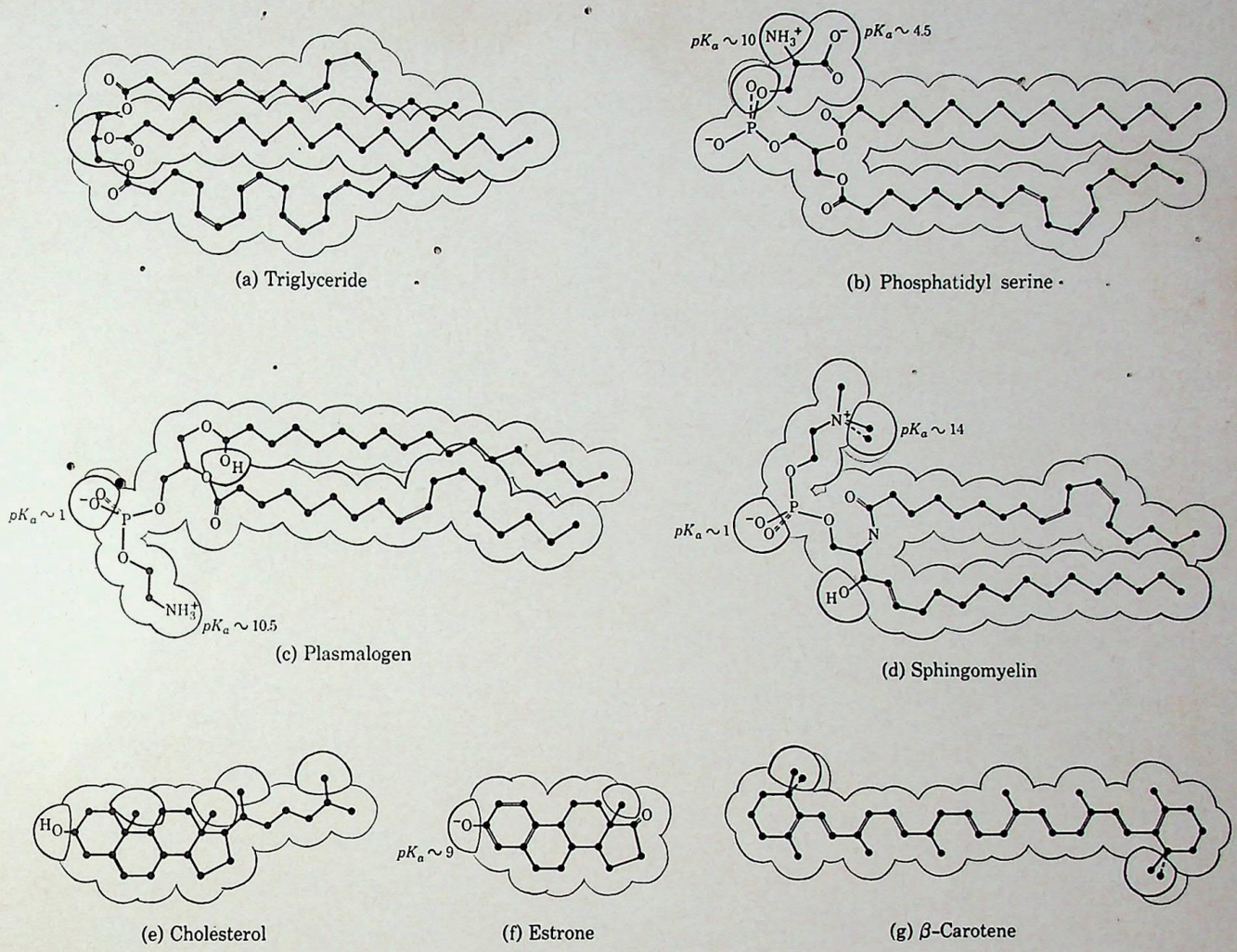


FIG. 17. Some selected lipid structures. The pK_a shown indicates roughly the acidity of the charged groups.

esters, etc.), there are lipid structures containing primary amine, tertiary amines, phosphoric-acid groups, phenolic hydroxyl groups, etc. It can easily be imagined that the lipids owe much of their importance in biological systems to the fact that they represent classes of materials which can bridge the gap from water soluble to water insoluble phases without necessitating a sharp discontinuity. They are often thought to be concentrated in biological membranes and interfaces. Lipids also provide the most compact form of storage of chemical energy.

The term "fatty acid" is not easy to define accurately, but is usually taken to include all of the straight-chain members of the acetic-acid series of carboxylic acids, and a number of naturally occurring unsaturated straight-chain acids (e.g., oleic, linoleic, arachidonic acids, etc.). The esters of these fatty acids include compounds with glycerol to form triglycerides, with α-glycerylphosphoric acid to form phosphatidic acids, with α-glyceryl phosphoryl esters to form lecithins, phosphatidyl ethanolamines, and phosphatidyl serines; with sphingosine to form sphingomyelin and cerebrosides; and with inositol N-phosphate to form phosphoinositides; etc. A few typical lipid structures involving certain of the lipid residues mentioned above are shown in Fig. 17. In (a) a typical triglyceride is shown, with arachidonic acid (4 double bonds), stearic acid (saturated), and linolenic acid (2 double bonds) joined to glycerol. The three fatty-acid residues are shown in their extended conformation, with the stearic-acid residue above the other two. The double bonds in the unsaturated fatty-acid residues are usually found in the *cis*-configuration. Phosphatidyl serine (b) represents a typical phospholipid, and is shown with one residue of linolenic acid and one of palmitic acid (saturated), joined through a glycerol residue to phosphoserine. Three ionizable groups are seen here—the strongly acidic phosphoric-acid residue (pK_a about 1), the weakly acidic carboxyl of serine (pK_a about 4.5), and the weakly basic amino residue of serine (pK_a about 10). In other phospholipids the serine residue may be replaced by ethanolamine ($HO-CH_2-CH_2-NH_3^+$) or by choline [$HO-CH_2-CH_2-N^+(CH_3)_3$]. In such phospholipids, no pK_a near 4.5 would be expected, and the basic group would be found to have a pK value of about 10 in the case of ethanolamine and above 14 with choline.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

The plasmalogens [Fig. 17(c)] are glycerophosphatides containing an active group (shown here as the —OH on carbon-1 of the stearic aldehyde residue) which reacts as an aldehyde group. The structure shown is one of several which have been proposed, and illustrates stearic-aldehyde and oleic-acid (1 double bond) residues joined to glycerol, which in turn is linked to phosphoethanolamine. This structure contains two ionizable residues—the phosphoric-acid group (pK_a about 1) and the amine group (pK_a about 10.5). Palmitic aldehyde sometimes replaces stearic aldehyde, and choline often replaces ethanolamine in typical plasmalogens. Other common types of lipids contain the lipid base sphingosine, $CH_3-(CH_2)_{12}-CH=CH-CH(OH)-CH(NH_2)-CH_2OH$. Illustrated here is a typical sphingomyelin [Fig. 17(d)] with a linolenic-acid residue joined through an amide linkage to the —NH— of sphingosine, and phosphocholine joined through an ester linkage to the terminal hydroxyl of sphingosine. This molecule would have ionizable groups in the phosphoric-acid and the choline residues (pK_a about 1 and above 14). The cerebrosides represent another typical sphingolipid type (not illustrated). Their structure is similar to sphingomyelin, but in place of phosphocholine a sugar residue, often D-galactose (as a pyranose ring), is joined to the terminal hydroxyl of sphingosine through a glycoside linkage. Such a cerebroside would not contain ionizable groups, but the many hydroxyl groups of galactose would serve as active sites for reaction with other molecules.

Cholesterol [Fig. 17(e)] is frequently found in animal tissues, either as the free alcohol (as illustrated here) or as the ester, in which the cholesterol hydroxyl is linked to a fatty acid. Neither cholesterol nor cholesterol esters contain ionizable groups, and the solubility properties of these molecules resemble those of the triglycerides. Estrone [Fig. 17(f)], a typical female sex hormone, has a ring structure much like cholesterol except that the first (*A*) ring contains three double bonds, the second ring is saturated, and the C-18 methyl is lacking. Estrone contains a ketonic oxygen at C-17, and a phenolic hydroxide residue (pK_a about 9) at C-3.

β-Carotene [Fig. 17(g)] is a highly unsaturated hydrocarbon, occurring in many plants and green leaves, and also in animal systems. It contains a highly conjugated system of eleven double bonds, responsible for the intense color and the reactivity of the hydrocarbon. An alcohol (vitamin A) and an aldehyde (vitamin A aldehyde) are of great importance in visual pigments and contain a hydrocarbon chain similar to one-half β-carotene. Two closely related hydrocarbons, α- and γ-carotene usually occur with β-carotene. The double bonds in the carotenoids usually are found to be in the *trans*-configuration (in contrast to the situation in the fatty-acid residues). The carotenoid hydrocarbons, along with other terpene-type compounds, appear to be built up from repeating isoprene units $[CH_2=C(CH_3)-CH=CH_2]$. Another hydrocarbon, squalene, $[CH_3-C(CH_3)=CH-CH_2-CH_2-C(CH_3)=CH-CH_2-CH_2-C(CH_3)=CH-CH_2-]_2$, is found in large amounts in shark-liver oil, and also is an intermediate for the synthesis of cholesterol in mammalian systems.

The mode of attachment of these lipid structures to the peptide moiety of a lipoprotein is poorly understood.[30] Denaturation of the protein and extraction with lipid solvents is usually sufficient to completely break the lipoprotein complex, so that it would seem that, if any covalent linkages are present, they must be very labile. A number of lipo-polysaccharides are known, and lipids are found to be firmly linked with some of the glycoproteins. It may be that the carbohydrate moiety of such complex macromolecules provides a covalent attachment for the lipid molecules. Much of the lipid material now thought to occur in cells as "free lipid" will, in the near future, be found to have intimate molecular binding to protein, glycoprotein, or carbohydrate cellular components.

ACKNOWLEDGMENTS

The author is greatly indebted to Dr. Margaret J. Hunter, Dr. Martha L. Ludwig, Dr. Donald F. H. Wallach, and Dr. Colin Green for their constructive criticism regarding this manuscript.

BIBLIOGRAPHY

[1] J. M. Bijvoet, A. F. Peerdeman, and A. J. van Bommel, Nature **168**, 271 (1950).
[2] B. W. Low in *The Proteins*, H. Neurath and K. Bailey, editors (Academic Press, Inc., New York, 1953), Vol. I., p. 235.
[3] L. Pauling in *Symposium on Protein Structure*, A. Neuberger, editor (Methuen and Company, Ltd., London; John Wiley and Sons, Inc., New York, 1958), p. 17.
[4] M. Calvin, Federation Proc. **13**, 697 (1954).
[5] G. E. Perlmann, Advances in Protein Chem. **10**, 1 (1955).
[6] D. F. Waugh, Advances in Protein Chem. **9**, 325 (1954).
[7] K. Linderstrøm-Lang in *Symposium on Protein Structure*, A. Neuberger, editor (Methuen and Company, Ltd., London; John Wiley and Sons, Inc., New York, 1958), p. 23.
[8] C. Tanford in *Symposium on Protein Structure*, A. Neuberger, editors (Methuen and Company, Ltd., London; John Wiley and Sons, Inc., New York, 1958), p. 35.
[9] C. Tanford, S. A. Swanson, and W. S. Shore, J. Am. Chem. Soc. **77**, 6414 (1955).
[10] J. Steinhardt and E. M. Zaiser, Advances in Protein Chem. **10**, 152 (1955).
[11] C. Tanford, J. D. Hauerstein, and D. G. Rands, J. Am. Chem. Soc. **77**, 6409 (1955).
[12] F. Sanger, Advances in Protein Chem. **7**, 1 (1952).
[13] F. Sanger, *Currents in Biochemical Research*, 1956, D. E. Green, editor (Interscience Publishers, Inc., New York, 1956).
[14] C. B. Anfinsen and R. R. Redfield, Advances in Protein Chem. **11**, 1 (1956).
[15] C. H. Li, Advances in Protein Chem. **11**, 101 (1956).
[16] *Ibid.*, **12**, 270 (1957).
[17] C. H. Li in *Symposium on Protein Structure*, A. Neuberger, editor (Methuen and Company, Ltd., London; John Wiley and Sons, Inc., New York, 1958), p. 302.
[18] J. I. Harris in *Symposium on Protein Structure*, A. Neuberger, editor (Methuen and Company, Ltd., London; John Wiley and Sons, Inc., New York, 1958), p. 333.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

[19] C. H. W. Hirs, W. H. Stein, and S. Moore in *Symposium on Protein Structure*, A. Neuberger, editor (Methuen and Company, Ltd., London; John Wiley and Sons, Inc., New York, 1958), p. 211.

[20] C. B. Anfinsen in *Symposium on Protein Structure*, A Neuberger, editor (Methuen and Company, Ltd., London; John Wiley and Sons, Inc., New York, 1958), p. 223.

[21] H. Tuppy in *Symposium on Protein Structure*, A. Neuberger, editor (Methuen and Company, Ltd., London; John Wiley and Sons, Inc., New York, 1958), p. 66.

[22] P. Jollés, J. Jollés-Thaureaux, and C. Fromageot in *Symposium on Protein Structure*, A. Neuberger, editor (Methuen and Company, Ltd., London; John Wiley and Sons, Inc., New York, 1958), p. 277.

[23] R. R. Porter in *Symposium on Protein Structure*, A. Neuberger, editor (Methuen and Company, Ltd., London; John Wiley and Sons, Inc., New York, 1958), p. 290.

[24] A. Neuberger, editor, *Symposium on Protein Structure* (Methuen and Company, Ltd., London; John Wiley and Sons, Inc., New York, 1958).

[25] W. Pigman, *The Carbohydrates* (Academic Press, Inc., New York, 1957).

[26] K. Meyer, Advances in Protein Chem. **2**, 249 (1945).

[27] G. E. W. Wolstenholme and M. O'Connor, editors, *Chemistry and Biology of Mucopolysaccharides, Ciba Symposium* (J. and A. Churchill, London, 1958).

[28] J. A. Lovern, *The Chemistry of Lipids of Biochemical Significance* (Methuen and Company, Ltd., London; John Wiley and Sons, Inc., New York, 1955).

[29] H. J. Deuel, Jr., *The Lipids* (Interscience Publishers, Inc., New York, 1951), Vols. I–III.

[30] E. Chargaff, Advances in Protein Chem. **1**, 1 (1944).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

5

Molecular Configuration of Synthetic and Biological Polymers

ALEXANDER RICH

Department of Biology, Massachusetts Institute of Technology, Cambridge 39, Massachusetts

IT is possible to make a rough division in biological systems between large molecules and small molecules. They may be regarded as having quite different roles. Small molecules are actively involved in metabolic cycles, and they provide most of the energy needed for the activity of a biological system. The large molecules often have quite a different role in that their configuration—the three-dimensional organization of the molecule—has a particularly important part in determining the function of the molecule. Indeed, on the molecular level, structure and function are closely related.

Almost all large molecules that are found in biological systems are polymers; that is, they are formed by the linear association of certain fundamental repeating units. This review discusses some of the general structural features of these polymeric molecules, starting with some synthetic polymers which illustrate certain features of structural organization, and then continuing on to a group of more biologically significant synthetic and naturally occurring polymers.

In recent years, there has been an increasing awareness of the widespread occurrence of helical polymer molecules, both from natural and synthetic sources. This is not surprising, since the helix is the most general configuration for a regular polymer. The position of adjacent residues along a helix can be specified by a translation distance and an angular rotation. The most familiar example of a helix is the twofold screw rotation in which adjoining residues are related by an angular rotation of 180°; however, this is a specialized configuration and is discussed in the following.

Most molecular structure determinations are the results of x-ray diffraction studies. For diffraction work on polymers, a very useful expression has been obtained by Cochran *et al.*,[1] who derived the Fourier transform for helical molecules. Their analysis utilizes the helical symmetry in the molecule and expresses the Fourier transform in terms of Bessel-function contributions to various layer lines. In this way, the position and intensity of the x-ray scattering by helical molecules can be computed relatively directly. Their analysis has been programed into computing machines in order to facilitate the testing of a variety of proposed helical molecular models.[2]

SYNTHETIC POLYMERS

Some understanding of the factors which determine the configuration of polymers can be gained from studies of the simpler synthetic polymers. It is illuminating to compare the configuration of polyethylene $-(CH_2)_n-$ with that of polytetrafluoroethylene $-(CF_2)_n-$ (known commercially as Teflon). The polyethylene molecule [Fig. 1(a)] consists of a series of CH_2 groups organized in a zigzag chain in which all of the carbon atoms lie in one plane. The distance between alternate CH_2 groups is 2.54 A. If, however, the hydrogen atoms are replaced by fluorine atoms, there is a change in configeration as is shown in Fig. 1(b).[3] Instead of lying in a planar zigzag, the carbon atoms in the backbone are now

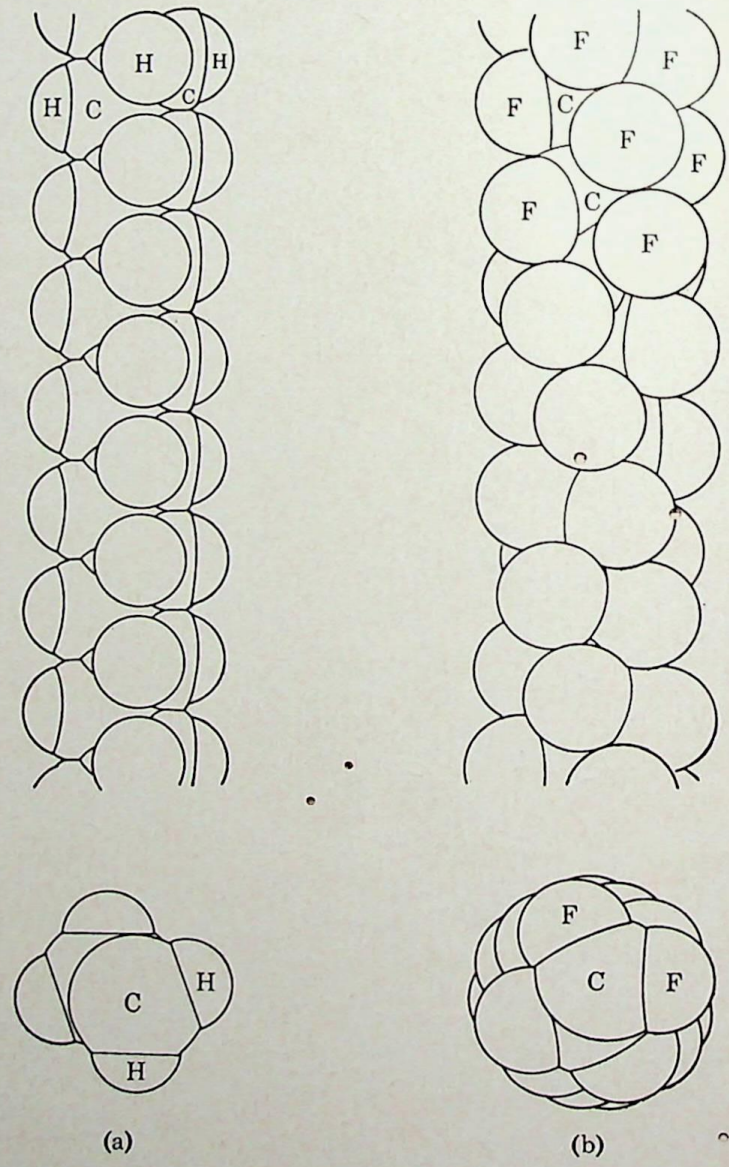


FIG. 1. (a) Planar zigzag chain formed by polyethylene (side and end views). The hydrogen atoms are not in van der Waals contact so that the carbon backbone remains fully extended. (b) Helical-chain configuration formed by the fluorocarbon molecule (side and end views). The bulkier fluorine atoms are in van der Waals contact and force the backbone chain of carbon atoms into a helical configuration [from C. L. Bunn and E. R. Howells, Nature 174, 549 (1954)].

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

twisted. In the hydrocarbon chain, adjacent residues are related by a rotation of 180°; in the fluorocarbon chain, this angle has changed to 166°, and the molecule has a helical appearance. The fluorocarbon chain assumes this form as a result of the steric restraints imposed on it by the bulkier fluorine atoms. The hydrogen atoms with a van der Waals radius of 1.1 A are so small that they are not in contact in polyethylene. However, the fluorine atoms with a van der Waals radius of 1.35 A are too bulky to lie in a planar zigzag, and a helical twist is introduced into the carbon backbone in order to fit each fluorine atom into the groove between the two fluorine atoms on adjacent residues. This fairly small structural change brings about a considerable alteration in the properties of the two materials. Looking down the axis of the fluorocarbon molecule [Fig. 1(b)], one can see that the polymer is rounded in cross section. This rounded contour is responsible for the first-order phase transition which occurs near 20°C, and is undoubtedly owing to the rotation of these molecules in the crystalline state about their molecular axes. The hydrocarbon molecule, with its flattened cross section, does not have a phase change of this sort. This is an example of the effect of atomic size on molecular configuration and chemical properties.

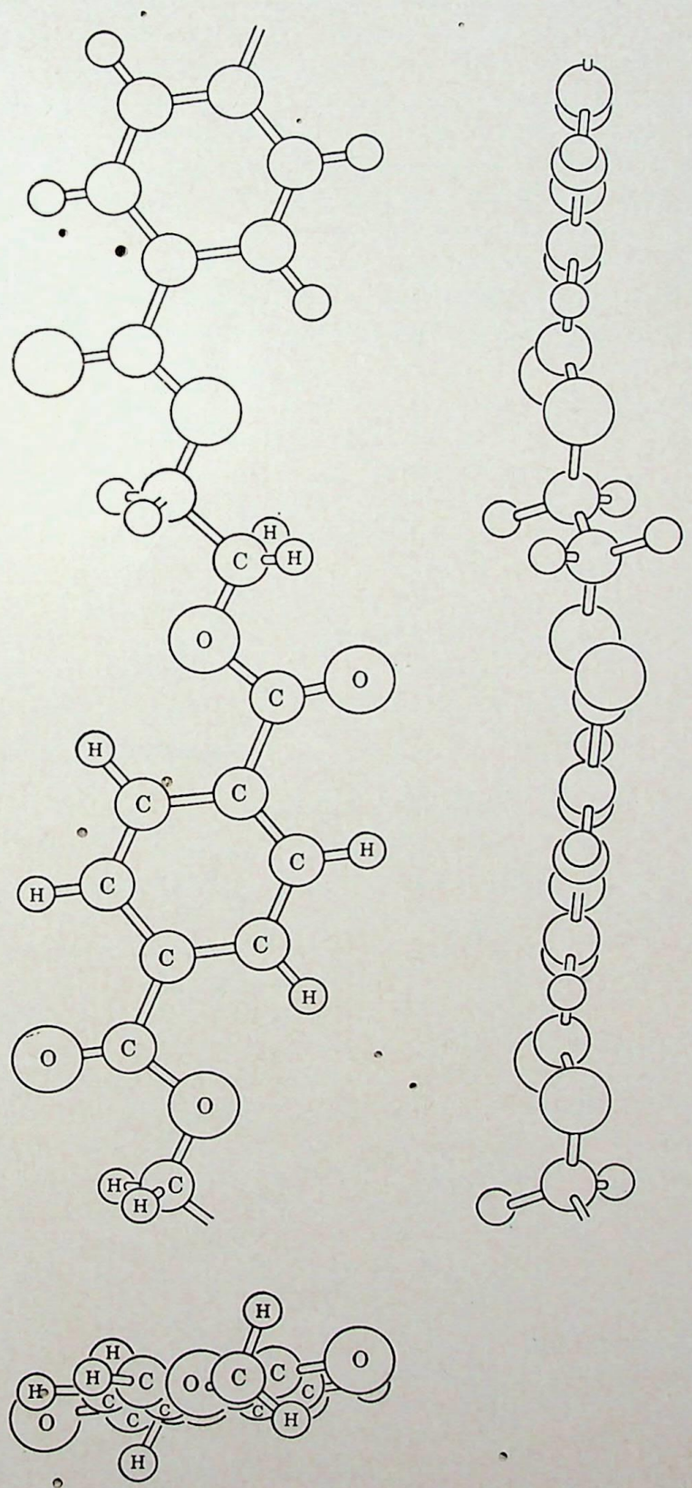


FIG. 2. Configuration of the polyethylene-terephthalate molecule. The unsaturated benzene groups lie nearly in the same plane so that the molecule has a ribbon-like configuration [from R. P. Daubeny, C. W. Bunn, and L. Brown, Proc. Roy. Soc. (London) **A226**, 531 (1954).

Another feature of interest is illustrated by the polymer molecule polyethylene terephthalate (known commercially as Dacron or Terylene). The repeating molecular unit contains a benzene ring as shown in Fig. 2. Succeeding residues lie almost in the same plane, hence the molecule has a ribbon-like configuration. Figure 3 illustrates how these molecules are packed in the crystal.[4] The flat, unsaturated rings lie largely on top of each other. In this orientation, they are stabilized considerably by the van der Waals interaction which is large for the highly polarizable π-electron system of the ring. This stabilization undoubtedly contributes to the high melting point (266°C) observed in the fiber. This type of interaction occurs in biological polymers and is most frequent in the nucleic acids where the large unsaturated ring systems usually pile on top of each other. These are discussed more thoroughly in the chapter devoted to the nucleic acids (p. 191).

When polymers contain electronegative atoms such as nitrogen and oxygen, there is often an opportunity for hydrogen bonds to be important in determining the configuration of the molecule. An example of this type of polymer is seen in polyhexamethylene adipamide, known as 66 nylon.[5] The structure of the α-crystal illustrated in Fig. 4 is most clearly understood by noting that the molecules form a sheet structure in which adjacent units on the "*a*" face of the unit cell are held together by the horizontal hydrogen bonds which are shown as dotted lines. This polymer is of particular interest, of course, because its residues are joined by peptide bonds (—NH—CO—) as are the amino-acid residues in a polypeptide or in a protein.

Hydrogen bonds represent fairly weak or secondary forces between atoms. Nonetheless, they are often of considerable importance in determining the configuration of polymeric systems, because the total number of bonds involved is of the same order of magnitude as the number of monomer units in the polymer molecule. This is usually a large number so that the total stabiliz-

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

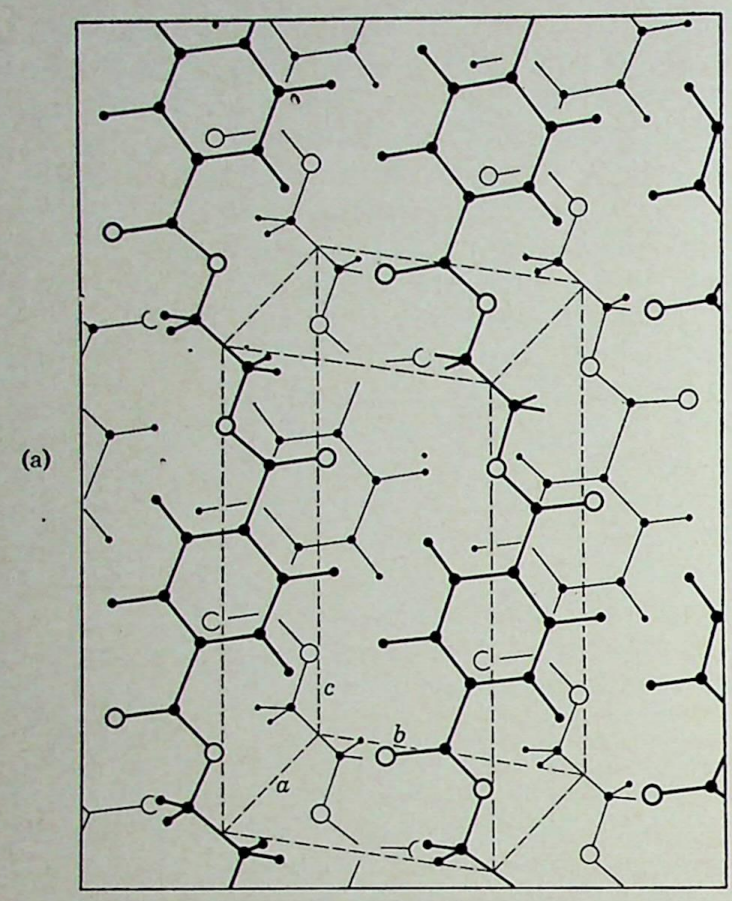


FIG. 3. The arrangement of polyethylene-terephthalate molecules in the crystal. (a) View perpendicular to the fiber axis. (b) View down the fiber axis. The larger dots represent carbon; smaller dots, hydrogen; open circles, oxygen [from R. P. Daubeny, C. W. Bunn, and L. Brown, Proc. Roy. Soc. (London) A226, 531 (1954)].

ing energy per covalently-linked molecule may become considerable.

BIOLOGICAL POLYMERS

In all of the polymer structures described in the foregoing, there is one regular repeating unit and the configuration of the molecule is the resultant of systematic and repeating steric effects, van der Waals stabilization, electrostatic forces, and hydrogen bonding. The biological polymers, however, have an additional element of subtlety. In the proteins, the peptide backbone is the repeating unit, while the sugar phosphate groups form the backbone in the nucleic acids. In addition to these repeating units, however, there is also a variety of side groups attached periodically in some definite sequence to the polymer backbone. Hence, in addition to the systematic, repeating interactions described in the foregoing, the configuration of the molecule must be dependent also upon the nature of these side groups and the sequence in which they occur.

In discussing the molecular configuration of biological polymers, one usually describes the results of structural investigations carried out in the solid state. This description is useful, since the solid-state structure often is retained in solution or in the living cell.

The biological polymers may be listed as follows:

(a) The nucleic-acid polymers. These are discussed at length in another chapter (Rich, p. 191). However, it should be mentioned that, in this polymer, each repeating residue has a negative charge located on the phosphate group. Thus, the configuration of a nucleic-acid molecule represents a compromise in which the internally repelling electrostatic forces are balanced by the cohesive forces in the molecule.

(b) The carbohydrate polymers. Only a few such as cellulose and chitin have a regular configuration with the polymer chains organized systematically. The ab-

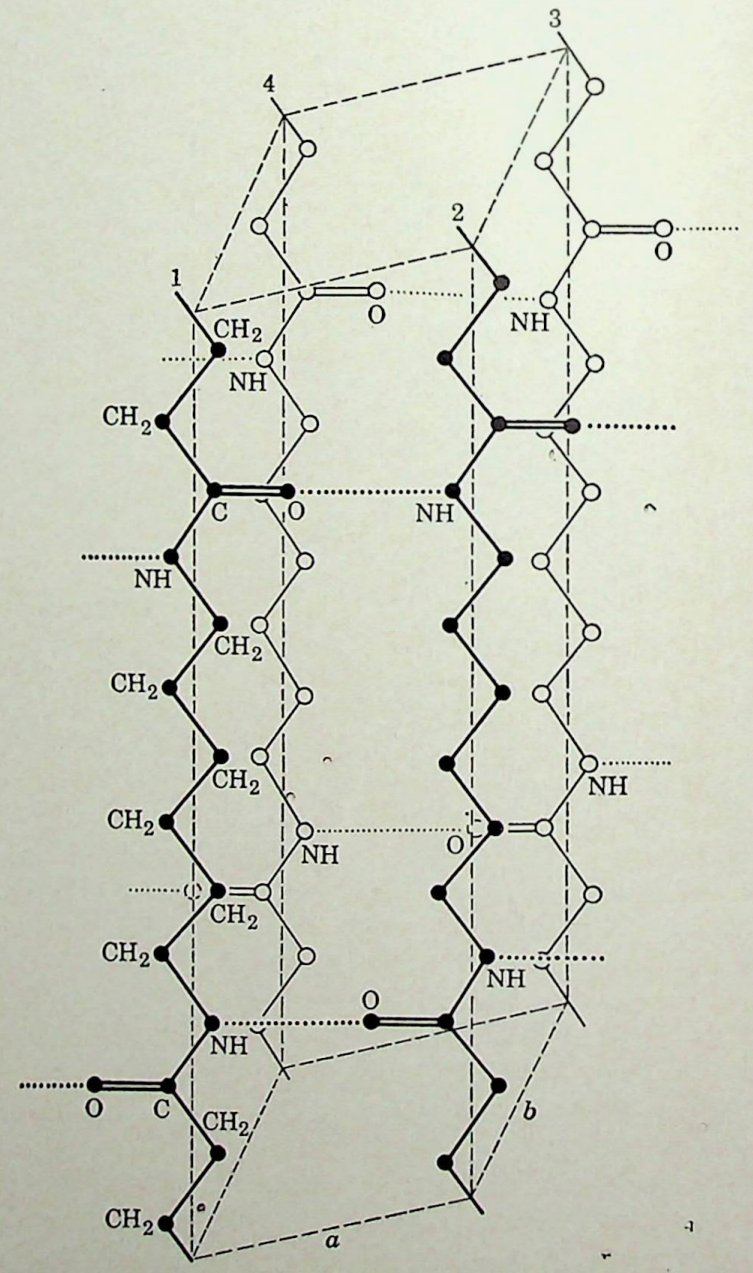


FIG. 4. The molecular structure of the α-crystal of 66 nylon. The molecules are hydrogen-bonded into sheets in the plane of the page [from C. W. Bunn and E. V. Garner, Proc. Roy. Soc. (London) A189, 39 (1947)].

©C-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

sence of a highly ordered configuration in most carbohydrate polymers is probably an indication that their biological function does not require any elaborate three-dimensional organization. Instead, it may depend on the stereochemistry of the sugar-to-sugar linkage. For example, the amorphous particles of glycogen stored in the cell are apparently just a convenient depot from which glucose residues can be removed by enzymatic action when they are needed. These structures are not discussed here.

(c) Proteins. Perhaps the most varied biological polymer is the polyamino-acid chain from which proteins are built. The repeating chemical unit is a single peptide bond with one α-carbon atom between each bond in contrast to the six carbon atoms between peptide bonds in 66 nylon. In the proteins, however, about twenty different types of side chains can be attached to the single α-carbon atom. The nature of these side chains has a profound effect in determining molecular configuration (see following).

(d) Lipids. There are no lipid polymers, even though there are several lipid molecules with extremely long side chains which are almost polymeric in nature.

It is interesting to note that there are apparently no mixed polymers in biological systems. Thus, one never finds a polymer chain involving both nucleotides and amino acids or sugar residues. If such mixed polymers existed, they might be unable to interact systematically through their backbones, and hence would lack an important element which gives rise to configurational stability.

POLYAMINO ACIDS

Present knowledge of the structure of the simpler polypeptides and proteins has its foundation in the extensive work of Pauling and Corey and their collaborators who worked out the crystal structures of a number of amino acids and small polypeptides. From their work, several important conclusions can be drawn.

(1) The dimensions of bond lengths and bond angles of the peptide group were obtained to an accuracy of ± 0.02 A and $\pm 2°$.[6–8] They found a similarity in the dimensions of the peptide group regardless of the particular amino acids involved in the linkage.[9] These are shown in Fig. 5. The carbon-oxygen distance in the carbonyl group was found to be 1.24 A, even though the sum of the double-bond covalent radii is 1.21 A. In a similar way, the carbon-nitrogen bond distance in the amide group is 1.32 A, which is less than the sum of the single-bond radii (1.47 A). This points to approximately 40% double-bond character in the C—N bond, and 60% double-bond character in the carboxyl group owing to resonance between the two.

(2) As a direct consequence of the double-bond character in the C—N bond, the amide group is expected to have a planar configuration. Experimental results amply confirm this expectation. It has been

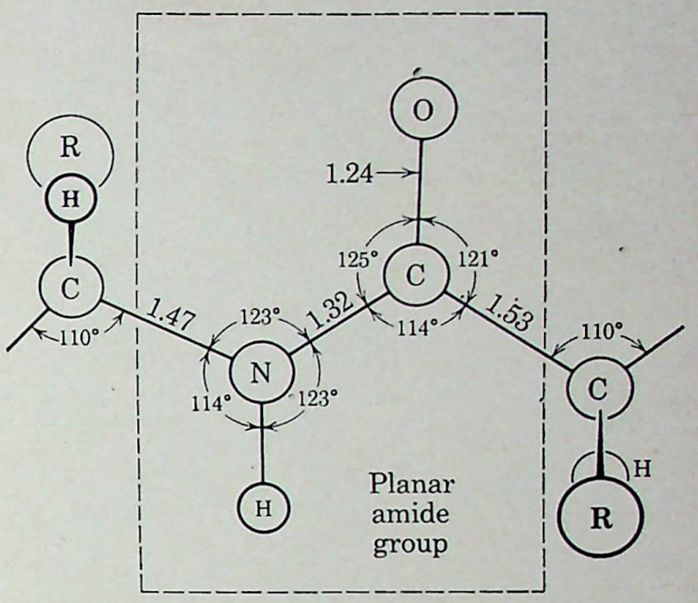


FIG. 5. Fundamental dimensions of polypeptide chains as derived from x-ray analyses of amino acids and simple peptides [from R. B. Corey and L. Pauling, Rend. ist. lombardo sci. 89, 10 (1955)].

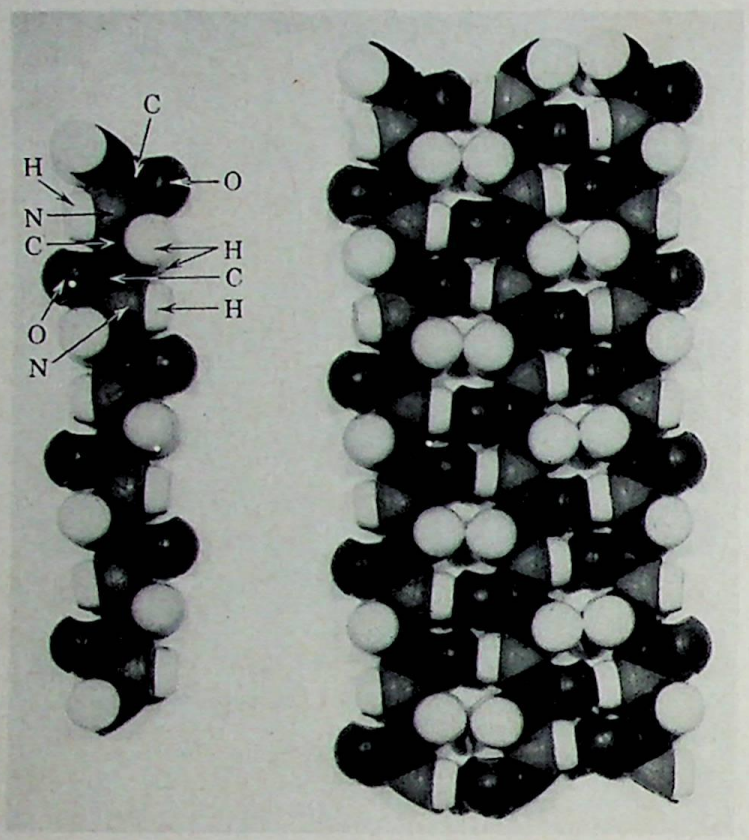
estimated that the distortion energy is about 1 kcal/mole for a deviation of 10° from planarity and about 3.5 kcal/mole for a 20° distortion.[6,7] In all of the peptide structures determined, the peptide group is found planar to within a few degrees.

(3) All of the small peptides now examined show a *trans*-configuration for the amide group. This is shown in Fig. 5 where the α-carbon atoms are attached to the planar amide group on opposite sides of the C—N bond. This conclusion from crystal-structure work is also supported by dipole moment studies.[10] Using infrared absorption, Badger and Rubaclava[11] have shown that the *trans*-configuration is more stable than the *cis*- by more than 2 kcal/mole.

(4) Another feature of the peptide group is the fact that it has a hydrogen atom attached to the nitrogen and is, therefore, capable of forming a hydrogen bond. All of the crystal structures analyzed were characterized by forming a maximum number of N—H···O hydrogen bonds. Most of the hydrogen bonds are linear to within a few degrees and have a length of 2.79 ± 0.12 A.[6,7]

Thus, the polypeptide backbone consists of a series of fairly rigid planar groups which may or may not be able to rotate relative to each other, depending upon the nature of the side groups attached to the α-carbon atom. This rotational barrier which may be as large as 3 kcal/mole is an additional important factor in determining the ultimate configuration of the molecules. Pauling and Corey first pointed out that a relatively small number of configurations will meet all of the four criteria described above. Thus, one can expect a fairly small number of stable polypeptide configurations. The following paragraphs review several of the more important ones.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA



(a) (b)

FIG. 6. A planar sheet composed of fully-extended polyglycine chains. The atoms are represented as solid out to their van der Waals radii. The hydrogen atoms attached to nitrogen have a depression in them for use in illustrating hydrogen bonding with carbonyl groups (>C=O). (a) Single extended polypeptide chain. (b) Antiparallel arrangement of polyglycine chains to form a sheet. Note that the hydrogen atoms in adjoining chains are in van der Waals contact. This would prevent the formation of this sheet from polypeptides other than polyglycine.

Extended Polypeptide Configurations

The simplest configuration of a polypeptide chain is one in which the polypeptide chain is extended so that all of the planar peptide units tend to lie in a common plane. In this extended form, alternate amino-acid residues have carboxyl groups pointing in the same direction, and adjacent residues are related by a translation of 3.6 A and a rotation of 180°. The usual extended form of polyglycine has a configuration which is close to that described in the foregoing. When fully extended, the polyglycine chain should have a repeat distance of 7.2 A. However, the fiber-axis repeat distance, obtained experimentally, is somewhat less[12]; hence, the chains are almost but not fully extended. An extended configuration is shown in Fig. 6, a photograph of atom models that are designed to show all of the van der Waals radii for the various atoms.[6] In Fig. 6, a single extended chain of polyglycine is shown also. This structure is helical in the same sense that polyethylene [Fig. 1(a)] is helical, with a rotation angle of 180°. Chains of polyglycine can be hydrogen bonded to make a sheet structure, as shown in Fig. 6(b), where the chains alternate in direction. Such a configuration in which the successive planar peptide residues are almost in the same plane could exist in polyglycine only and not in any other polypeptide, because steric interference would result in the presence of a bulkier side chain. This can be seen most clearly by noting that the hydrogen atoms in the sheet structure are in contact. Thus, as illustrated in Fig. 6, it would be impossible to have a larger β-carbon atom attached in the position of the hydrogen atoms.

It is possible, however, to make a sheet structure utilizing hydrogen bonds in the plane of the sheet by rotating the peptide units until the β-carbon atoms are no longer in contact as they are in the polyglycine configuration of Fig. 6. This configuration is important in the structure of silk fibroin.

Structure of Silk Fibroin

There is a rotational barrier between adjacent planar peptide units in the polypeptide chain, and, as a consequence, only a limited number of configurations are stable. One of the earliest to be investigated critically by Pauling and Corey was a semi-extended form in which adjacent planar peptide groups were related to each other by a translation of approximately 3.5 A and a rotation of 180°. A drawing of the configuration, looking down the plane of the peptide units, is illustrated in Fig. 7(a). In this configuration, the β-carbon atoms emerge from alternate sides of the polypeptide backbone, and the backbone itself has a zigzag rather than a fully extended form. The most significant feature of this configuration is the orientation of the C=O and N−H groups which line in one plane and are perpendicular to Fig. 7(a). These groups are oriented so that all hydrogen bonds can be formed by aligning a series of these polypeptide chains side by side. This has been done in Fig. 7(b) to form what is called the "antiparallel chain pleated sheet" structure. Alternate chains run in opposite directions, and in this way the successive amino and carboxyl groups are hydrogen-bonded to chains on either side of the original chain. By working carefully with molecular models, it is possible to show that the identity period along the fiber axis is 7.00 A, whereas the lateral distance between equivalent chains (alternate chains) is 9.50 A.

There are several interesting features in the antiparallel chain pleated sheet structure. The β-carbon atoms protrude at right angles from the plane of the sheet on alternate sides for successive residues along one chain. Furthermore, the β-carbon atoms are in phase on either side of the sheet; that is, they all appear at the same points along the fiber axis. Thus, there are large grooves between adjacent rows of β-carbon atoms.

The common form of silk fibroin is obtained from the silkworm *Bombyx mori*, and it has been the subject of x-ray diffraction investigations for many years. Most recently, Marsh *et al.*[13] have carried out an extensive and careful investigation and have arrived at a structure for this form of silk fibroin based on the antiparallel chain pleated sheet structure. Although the diffraction pattern is not simple, most of the reflections can be

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

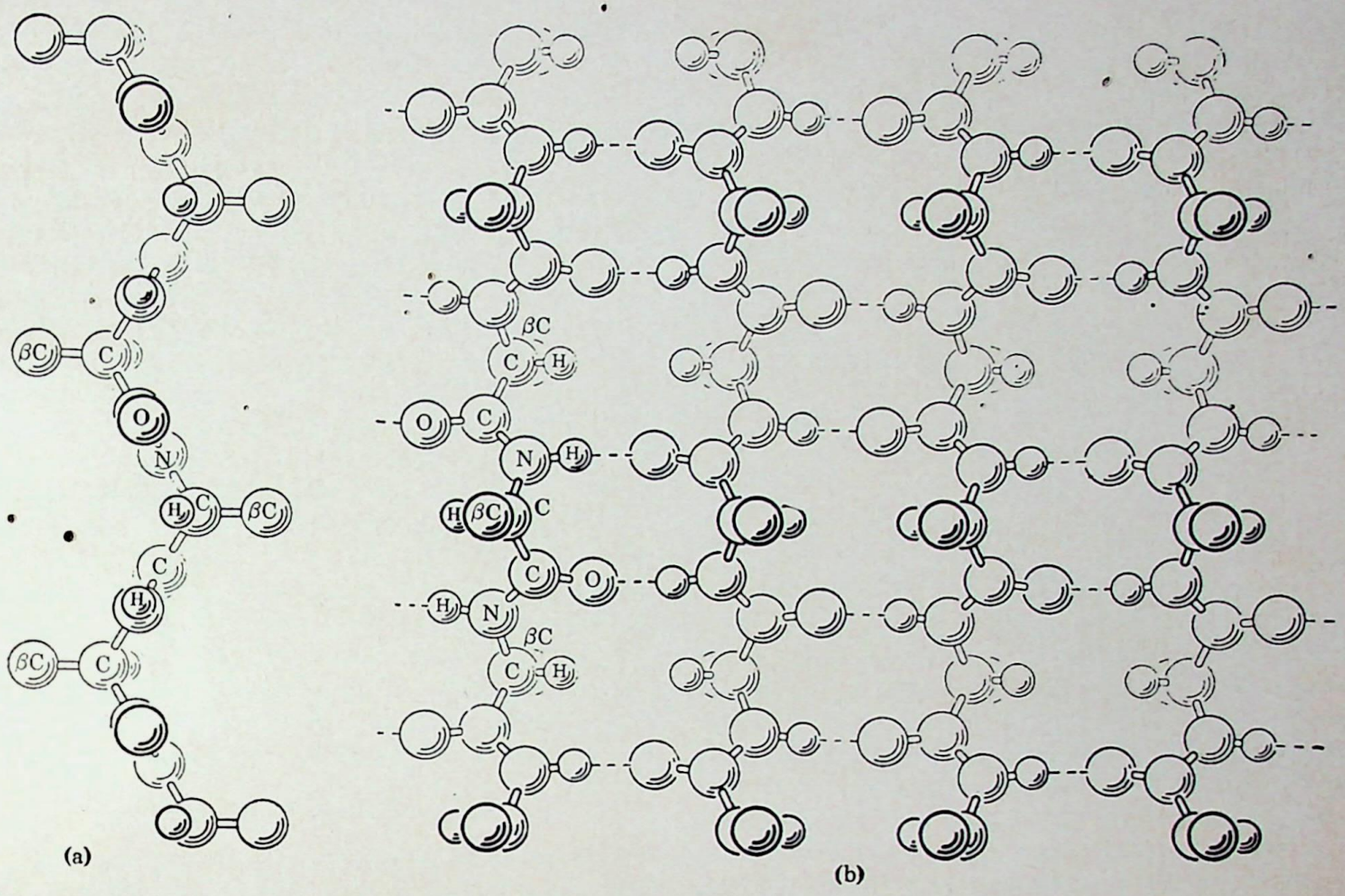


FIG. 7. (a) A drawing of a single polypeptide chain in the configuration found in the antiparallel chain pleated sheet. The chain is viewed in the direction of the hydrogen bonds and, therefore, along the plane of the sheet. Note the zigzag arrangement produced by alternate planar peptide residues. (b) A drawing of the antiparallel chain pleated sheet. In a given polypeptide chain, alternate β-carbon atoms are found on the same side of the pleated sheet [from R. E. Marsh, R. B. Corey, and L. Pauling, Biochim. et Biophys. Acta **16**, 1 (1955)].

accounted for by a simple unit cell which is orthogonal with dimensions 9.20 and 9.40 A at right angles to the fiber and a fiber-axis repeat of 6.97 A. The similarity between two of these dimensions observed experimentally and the pleated-sheet dimensions (7.00 A along the polypeptide chain and 9.50 A between equivalent alternate chains) suggested to these investigators that the antiparallel chain pleated sheet structure might serve as the structural basis of silk fibroin. Experiments carried out with rolled and flattened samples of silk fibroin produce the double orientation usually associated with layered structures. In this way, they were able to show that the 9.20-A reflection represented the repeat distance between molecular layers in silk fibroin. To deduce the structure, they had to arrange the layers of the sheet structure such that an identity repeat of 9.2 A was obtained.

One of the most remarkable features of silk fibroin is its unusual chemical composition. The simplest of the amino acids, glycine, constitutes almost one-half (44%) of the total number of amino acids present, while the next simplest amino acids (alanine and serine) comprise almost 40% of the residues. This suggested to the investigators that a simplified model of the structure might contain glycine residues with the alanine and serine residues in between at alternate sites along the polypeptide chain. In this way, all of the side groups protruding from one side of the sheet are from glycine residues and, therefore, consist simply of hydrogen atoms. The other alternate sites are occupied by other side chains—mainly alanine which has one methyl group or serine which has CH_2OH as a side chain. When the

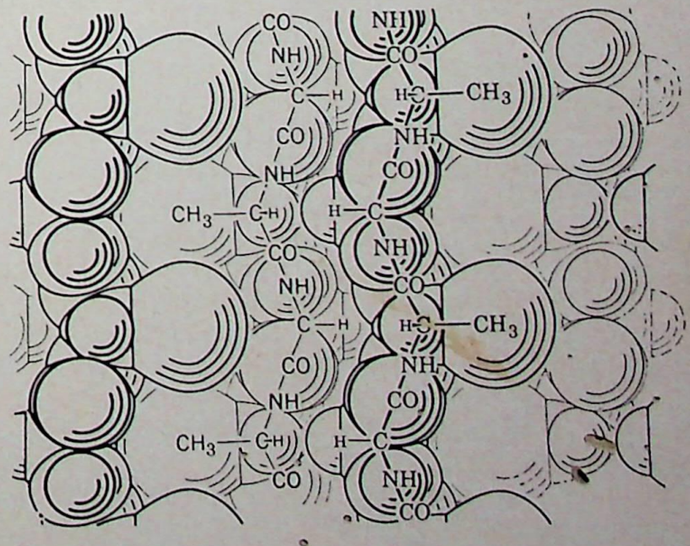


FIG. 8. A packing drawing of the structure of silk fibroin (*Bombyx mori*) viewed parallel to the pleated sheet. The fiber axis is vertical in this view [from R. E. Marsh, R. B. Corey, and L. Pauling, Biochim. et Biophys. Acta **16**, 1 (1955)].

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

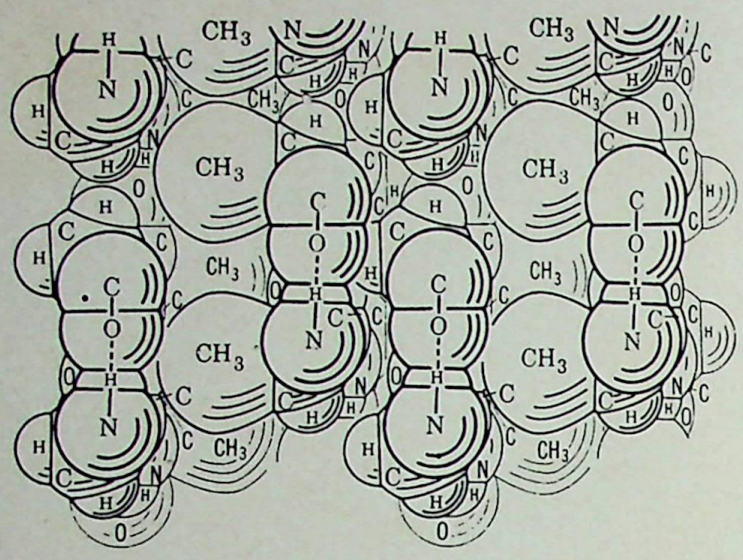


FIG. 9. A packing drawing of the structure of silk fibroin (*Bombyx mori*) viewed along the fiber axis. The hydrogen-bonding sheets are vertical in this view [from R. E. Marsh, R. B. Corey, and L. Pauling, Biochim. et Biophys. Acta **16**, 1 (1955)].

side chains are distributed in this manner, it is possible to pack pairs of sheets together so that the bulky side chains are located between the two sheets. In this way, the β-carbon atoms of one sheet protrude into the space between the β-carbon atoms from the other sheet. The distance from one sheet to the next is 5.7 A when they are packed in this way. However, when these double-sheet structures are stacked together, the glycine side chains are in contact and the distance from sheet to sheet over the glycine contacts is 3.5 A. Hence, the identity repeat for such a head-to-head, tail-to-tail stacking of sheets is $5.7+3.5=9.2$ A. This is, of course, exactly what is observed in the simple unit cell derived from the x-ray diffraction study of silk fibroin.

A packing drawing of the structure of silk fibroin is shown in Figs. 8 and 9. In Fig. 8, the view is perpendicular to the fiber axis and parallel to the pleated sheets. This view shows how the bulky methyl side chains are packed together with adjacent sheets filling alternate positions along the fiber axis, and how the small hydrogen side chains of the glycine residues fit together. Figure 9 is a view down the fiber axis. In this view, the hydrogen-bonding between the chains can be seen in the vertical direction, and the alternate packing arrangement of side chains is also clearly shown.

The reflections calculated from a simplified structural model of this type agree very well with the intensities obtained from the x-ray diffraction study of silk fibroin. This model does not account for all of the amino acids found in silk, including many with bulky side chains. They act to perturb this simplified model and produce some of the minor complications present in the diffraction pattern. Most of the properties of silk, however, can be understood from the simplified unit of structure.

Coiled Helical Configurations

The extended form of polyglycine shown in Fig. 6 is called polyglycine I and was the first form studied crystallographically by x-ray methods. In 1955, Bamford and his colleagues[14] described a second form of polyglycine which they called polyglycine II. The struc-

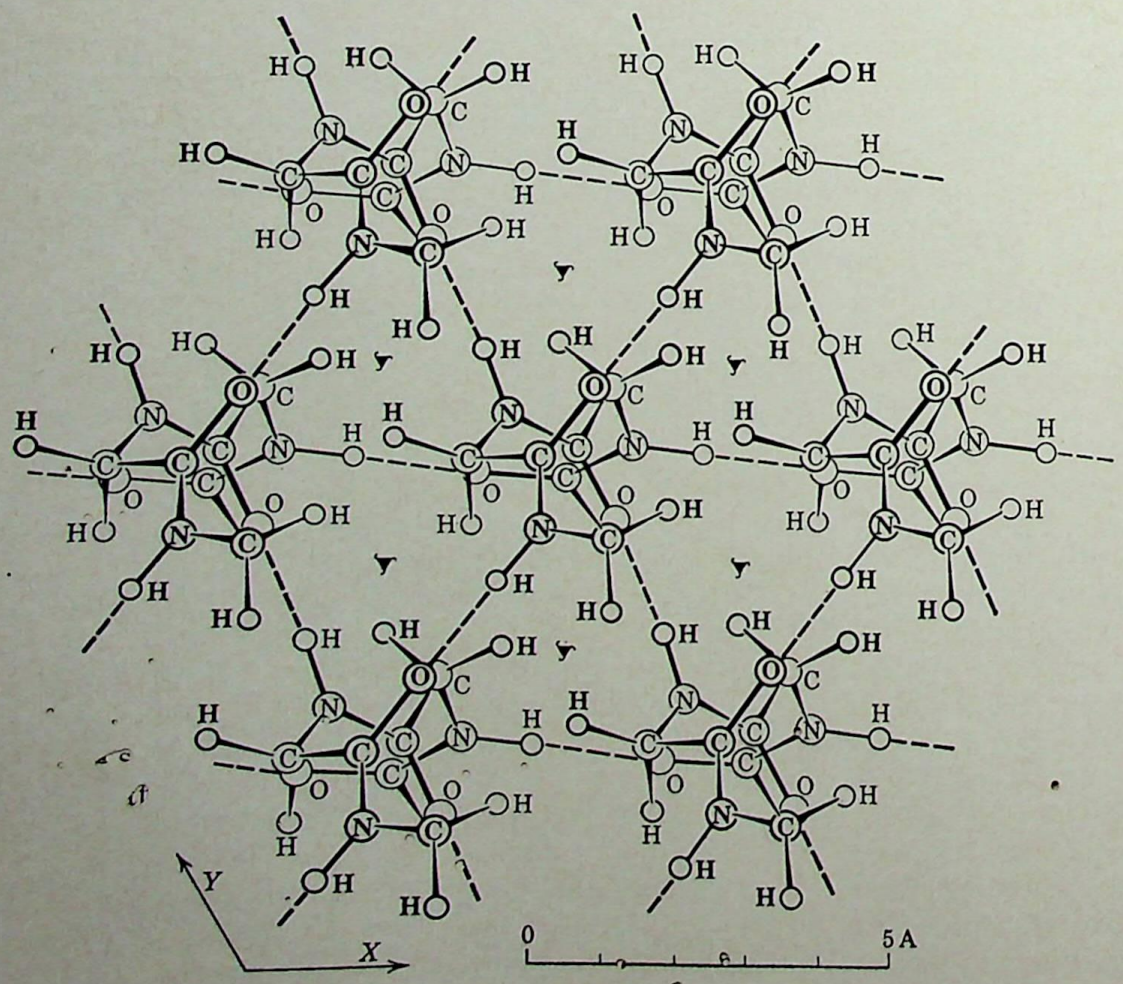
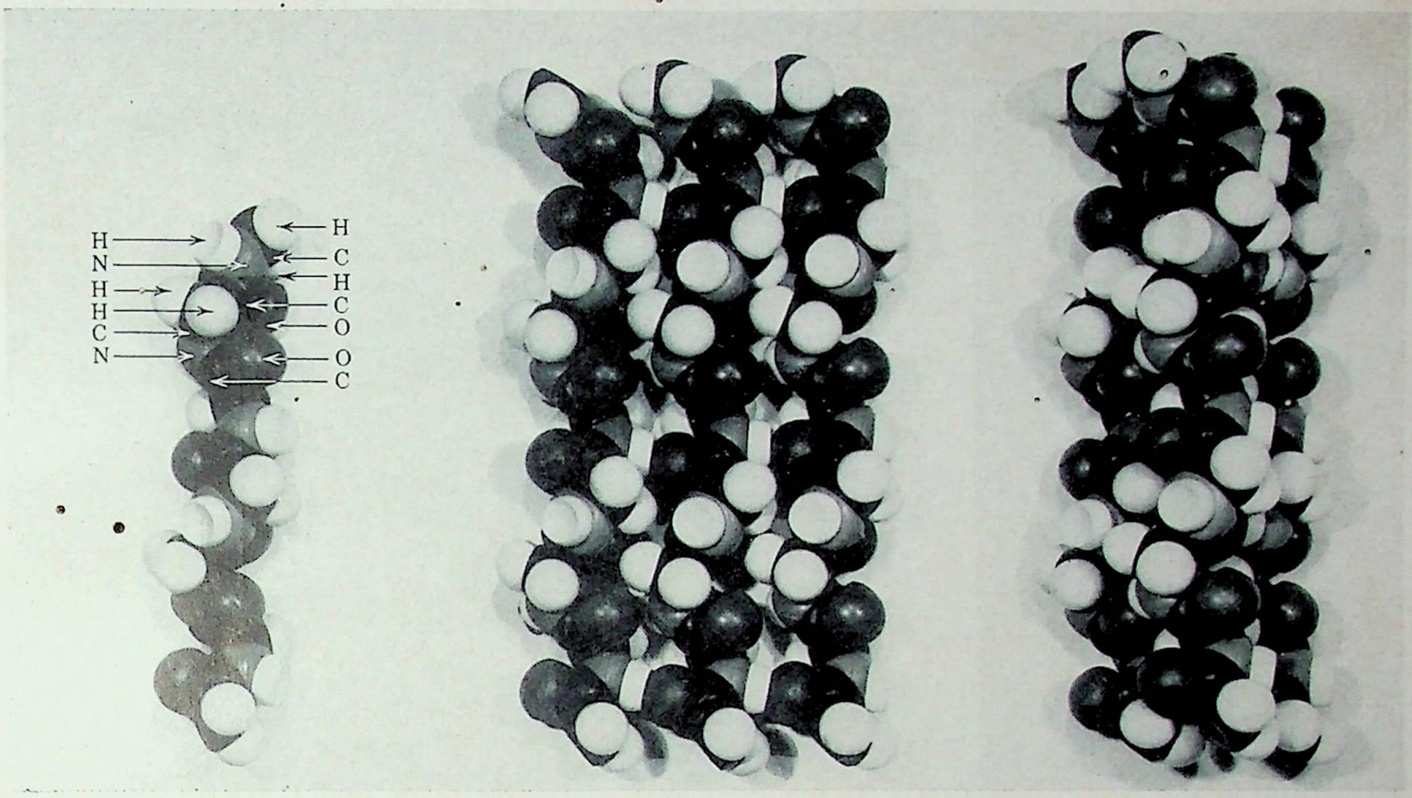


FIG. 10. The crystal structure of polyglycine II as viewed along the axis of the helical chains. The dashed lines represent hydrogen bonds [from F. H. C. Crick and A. Rich, Nature **176**, 780 (1955)].

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA



(a) (b) (c)

FIG. 11. (a) Molecular model of a single polyglycine chain arranged in the helical form found in polyglycine II. There are three peptide residues per turn in this form. (b) Three chains hydrogen bonded together to make a planar part of the polyglycine-II lattice. (c) Three chains packed in triangular form as found in the polyglycine-II lattice. The hydrogen bonds connecting these three chains are partly obscured.

ture of polyglycine II has been worked out by Crick and Rich,[15] and it is a particularly simple one. In it, all of the polypeptide chains are parallel, and each one is organized into a threefold screw axis so that adjacent residues are related by a translation of 3.1 A and rotation of 120°. The chains are packed in an hexagonal array, each chain being hydrogen bonded to six neighbors. The hydrogen bonds lie roughly perpendicular to the screw axis, and run in several directions and not solely in one plane as in the polyglycine sheet structure described in the foregoing. Figure 10 is a view of this structure down the fiber axis. In it, seven polypeptide chains are hexagonally packed with dashed lines indicating the linear hydrogen bonds which hold the adjoining chains together. The hydrogen-bond distance is 2.76 A, and the spacing between the polypeptide chains is 4.8 A. Since the translation along the axis is 3.1 A per residue, the axial repeat is 9.3 A because of the threefold screw symmetry. In this configuration, the polypeptide chains can fulfill all of the criteria developed by Pauling for stable configurations of polypeptide chains. It should be noted, however, that the polyglycine-II lattice could be formed only by a polymer of glycine. If bulkier side chains were present, the molecules would not be able to hydrogen bond, since the chains are already densely packed with only hydrogen atoms attached to the α-carbon atoms. A side view of the polyglycine-II chain structure is seen in Fig. 11, using molecular models which show the atomic van der Waals distances. On the left [Fig. 11(a)] is a single polyglycine chain arranged in a threefold screw axis. In the center [Fig. 11(b)], three of these chains are hydrogen-bonded to make one plane through the polyglycine-II lattice. It can be seen that only one-third of the hydrogen bonds is utilized in forming a sheet. The remaining two-thirds of the hydrogen bonds are used in holding the sheets together. Thus, in order to form all of the hydrogen bonds, a three-dimensional network must be built up in polyglycine II in contrast to the two-dimensional sheet structure seen in polyglycine I. In Fig. 11(c), the same three polyglycine chains are present as are in Fig. 11(b), but they are arranged so that they are no longer planar. One polyglycine chain is lying on top of the other two so that the three are hydrogen bonded. In this form, they represent a group of three polyglycine chains obtained by selecting a triangle of three close-packed chains (see Fig. 10). It should be pointed out that there are two ways of selecting a group of these three polyglycine chains. This can be seen most readily by observing the directions of the hydrogen bonds (N—H···O) which hold together the group of three. In one group, they are clockwise and, in the other, they are counterclockwise. This difference in the two groups is a consequence of the space-group symmetry, since the chains are arranged about a threefold screw axis and not about a sixfold screw axis.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

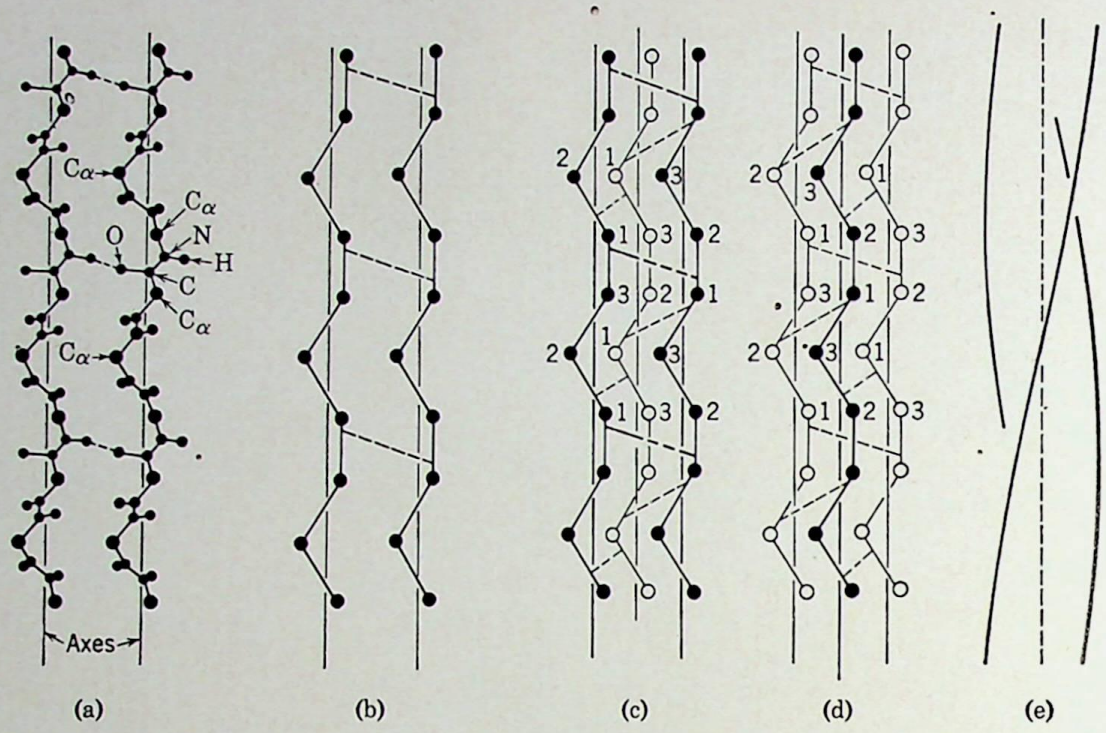


FIG. 12. A schematic drawing which illustrates how the collagen molecule may be generated from the polyglycine-II lattice. (a) Two molecular chains from polyglycine II. (b) The same chains showing only the α-carbon atoms (solid circles) and hydrogen bonds (dashed lines). (c) A third chain is added behind the two chains of (b). This generates collagen I. (d) The third chain is added in front of the two chains found in (b). This generates collagen II. (e) In the collagen molecule itself, the three molecular axes are coiled as shown in the diagram. The numbers on (c) and (d) correspond to the three kinds of sites found on the collagen molecule. Site number 1 is usually occupied by glycine [from A. Rich and F. H. C. Crick in *Recent Advances in Gelatine and Glue Research* (Pergamon Press, London, 1957), p. 1].

Structure of Collagen

Just as the structure of the antiparallel pleated sheet serves as a model for deriving the structure of silk fibroin, the polyglycine-II lattice can be used to derive the structure of collagen. Collagen, found in the form of elongate fibrils, is the major structural element of skin and connective tissue in the animal kingdom. One of its most unusual features is the chemical composition. One-third of its residues is glycine, and it has, in addition, a large number of pyrrolidine rings. Proline plus hydroxyproline make up about 24% of the residues. Because of the high glycine content, it is perhaps not surprising that there is a relationship between the polyglycine-II lattice and the collagen molecule. This relationship can be seen most clearly in Fig. 12, which shows a schematic development starting from the polyglycine-II lattice and ending in the collagen molecule.[16] In Fig. 12(a), two parallel polyglycine chains are sketched with the hydrogen bonds joining them (dotted lines). In Fig. 12(b), only the α-carbon atoms of the glycine residues are shown, with short solid lines schematically indicating the peptide units connecting them. It is seen that these polypeptide chains coil around the two axes and are held together by the hydrogen bonds (dashed lines). The collagen molecule may be derived from three polyglycine chains such as is shown in Figs. 12(c) and 12(d). In Fig. 12(c), a third polypeptide chain is added behind the two shown in Fig. 12(b), whereas in Fig. 12(d), the third chain is added in front of the original two. In this way, two groups of three polypeptide chains are created which are prototypes of two models, collagen I and collagen II. These correspond to the two different ways of collecting groups of three molecules as shown in the end view of the polyglycine-II lattice (Fig. 10). In the collagen molecule, however, the axes around which the polypeptide chains coil are no longer straight as in Figs. 10(a)–10(d), but are coiled slightly as shown in Fig. 12(e). Thus, the collagen molecule is regarded as a coiled-coil structure.

In silk fibroin, the spacings between the sheets in the molecule are uneven; the closer spacing (3.5 A) arises because there are only the side chains of glycine between the sheets. Similarly, in the collagen molecule, the three polypeptide chains can come close together, because the glycine residues are located near the center of the group of three chains, whereas the other two-thirds of the residues are projecting outward from the center of the molecule. Thus, two-thirds of the residues can have any side chains on them. And, what is more significant, they can accommodate pyrrolidine rings without producing any steric difficulties.

It is generally agreed that the collagen-II model is the most likely form for the collagen molecule,[17–19] and in it pyrrolidine rings can be found on two of the externally situated sites, while the third internal site must have a glycine residue. This explains why all collagens are found with 33% glycine in their composition. Figure 13 shows a form of the collagen-II molecule where two-thirds of the sites are filled with pyrrolidine rings such that a repeating sequence is used in each chain. In it, the sequence glycine-proline-hydroxyproline is used, since this is a common constituent of collagen

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

digests. This model of collagen was made using the polypeptide chains of Fig. 11(c), except that the axes of the polyglycine chains are now coiled rather than straight. As seen in Fig. 13, the collagen-molecule model is elongated, and has a distinct helical groove running down it which arises from the absence of a bulky side chain on the glycine site. The helical ridge on the surface of the molecule arises because the bulky side chains, in this case pyrrolidine rings, all lie closely packed. In the collagen molecule, however, there is a variety of amino acids found in that ridge, including many which have polar side chains. These are undoubtedly important in stabilizing collagen by hydrogen bonding, as well as in forming electrostatic bonds from one molecule to its nearest neighbor.

In the polyglycine-II lattice, the fundamental screw operation used in generating the lattice is a translation of 3.1 A and a rotation of 120°. Because collagen is a coiled-coil molecule formed from this lattice, the fundamental screw operation in generating the collagen helix is a translation of 2.86 A and a rotation of 108°. Of course, the asymmetric unit in the collagen molecule now consists of a group of three amino acids in contrast to the single amino acid in the polyglycine chain. In collagen, the coiling of the three polypeptide chains about one another is produced by the steric interaction of the β-carbon atoms which are present in the collagen molecule but which are absent in the polyglycine-II lattice.

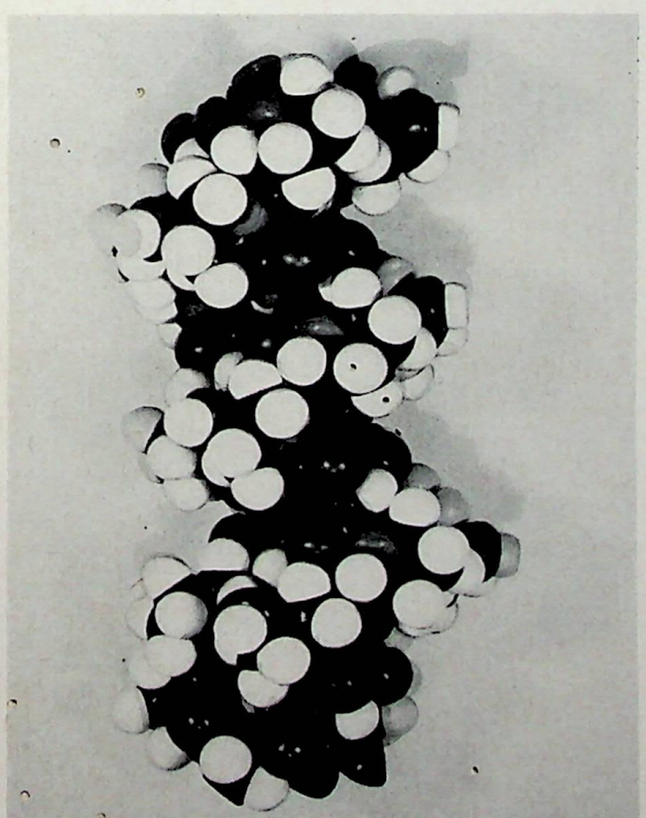


FIG. 13. A model of collagen II. The three polypeptide chains of Fig. 11(c) have been modified by the addition of proline and hydroxyproline residues to sites 2 and 3. This causes the three chains to coil about as shown in Fig. 12(e).

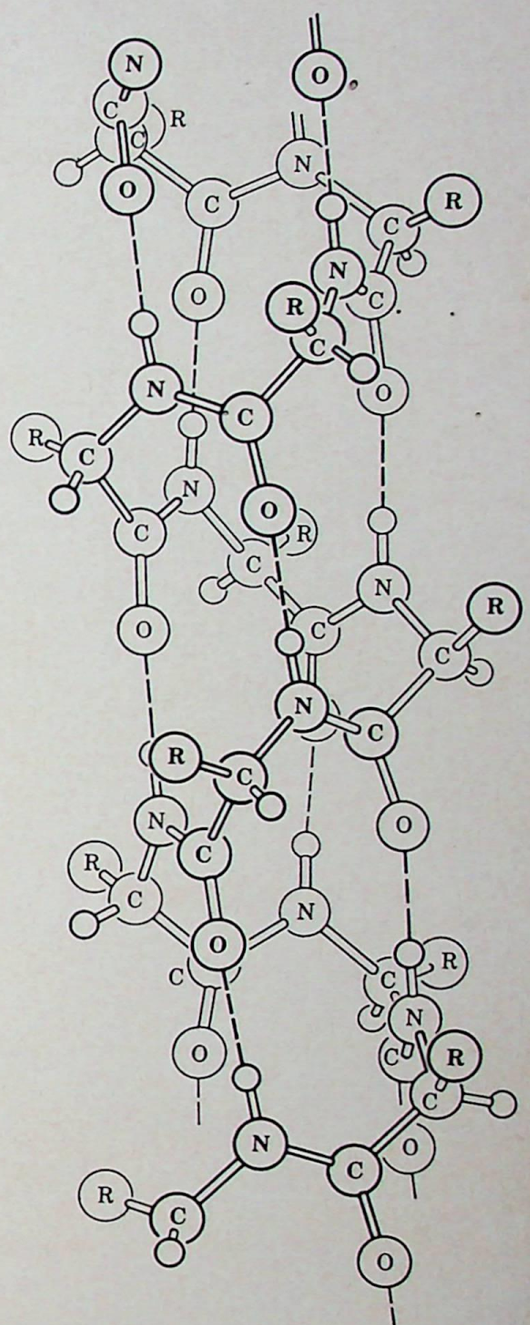


FIG. 14. A drawing of the α-helix [from R. B. Corey and L. Pauling, Rend. ist. lombardo sci. 89, 10 (1955).

The α-Helix

An important feature of the structure of silk fibroin is the fact that alternate residues are glycine units along the polypeptide chain. This permits the sheets to pack very closely. Similarly, in the collagen molecule, an important feature is the fact that every third residue is glycine, thereby allowing the three chains to come close together so that they can hydrogen bond. In a sense, both of these structures are very specialized and are a consequence of a restricted amino-acid composition. The best example of a polypeptide configuration which can be formed with minimal or no compositional restrictions is the α-helix.[20] In the α-helix (Fig. 14), the peptide units are oriented so that their planes are tangent to a cylinder

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

around the axis of the helix. In this way, hydrogen bonds can be formed which are parallel to the axis, and thus serve to hold together the various turns of the helix. The fundamental operations for generating the α-helix are a rotation of 100° around the axis and a translation of 1.5 A along the axis. Each amide group is connected by a hydrogen bond to the third amide group from it along the peptide chain. There are 3.6 residues per turn of the helix, and the total rise of the helix per turn is 5.4 A. The helix formed in this way is packed very firmly, and there is no open space in the center.

Of all of the configurations discussed hereto, the α-helix is by far the most important. It is a completely general structure in that it will accommodate any amino-acid side chain. When proline residues are incorporated into the structure, the helix axis changes direction, but this incorporation probably does not destroy the organization of the helix. This structure has been found in a variety of synthetic polypeptides,[21,22] and there is evidence for its existence in many globular proteins as well as hair, muscle, wool, and myosin.[23,24] The α-helix should produce a reflection at a spacing of 1.5 A, which is the translation distance along the fiber axis for each amino-acid residue. Perutz[23] has found this reflection in a variety of proteins, and this has provided strong inferential evidence for the α-helix.

It has been suggested that the α-keratin proteins such as hair contain helices twisted around one another to form coiled-coil structures.[25,26] Thus, these molecules may form by hexagonal packing into cable-like structures involving several strands which coil around each other. Structures of this sort give reasonable agreement with the observed x-ray diffraction photograph obtained from hair.

The importance of the α-helix lies in its extreme generality. Knowledge of the detailed configuration of globular proteins, as discussed by Kendrew (p. 94), is just beginning to be advanced. It seems quite likely that the α-helix will form an important feature of many of these molecules. Significantly, the α-helix buries the systematically repeating part of the polypeptide chain, using it to form a supporting fabric of great stability. In so doing, it exposes to the external environment all of the side chains with their great diversity. Thus, the α-helix enables a protein structure to maximize its variability and, thereby, produce the considerable chemical versatility of the proteins.

SUMMARY

On considering the molecular structure of biological polymers, the proteins and the nucleic acids, one can see that they are more complex than are synthetic polymers. Nonetheless, there are many common structural features in all of these materials. The helix is seen as the most general form for organizing successive residues along a polymer chain. The helix is found in many forms, ranging from fully extended to tightly coiled molecules. In all of these structures, space is filled efficiently, thereby maximizing the amount of van der Waals stabilization. Most synthetic polymers achieve a limited degree of structural complexity, as for example, coiling into a helical form. Materials of biological origin, by contrast, often go beyond this in complexity. They may form coiled-coils, as in collagen or in the α-keratin structure. On even a higher level, these polymers may fold in a very complicated fashion to produce globular molecules in which the interactions of the different side chains play a critical part in maintaining configurational stability.

BIBLIOGRAPHY

[1] W. Cochran, F. H. C. Crick, and V. Vand, Acta Cryst. **5**, 581 (1952).
[2] D. R. Davies and A. Rich, Acta Cryst. (to be published).
[3] C. W. Bunn and E. R. Howells, Nature **174**, 549 (1954).
[4] R. P. Daubeny, C. W. Bunn, and C. J. Brown, Proc. Roy. Soc. (London) **A226**, 531 (1954).
[5] C. W. Bunn and E. V. Garner, Proc. Roy. Soc. (London) **A189**, 39 (1947).
[6] R. B. Corey and L. Pauling, Proc. Roy. Soc. (London) **B141**, 10 (1953).
[7] R. B. Corey and L. Pauling, Rev. Sci. Instr. **24**, 621 (1953).
[8] L. Pauling and R. B. Corey, Fortschr. Chem. org. Naturstoffe **8**, 310 (1954).
[9] R. B. Corey and L. Pauling, Rend. ist. lombardo sci. **89**, 10 (1955).
[10] S. Mizushima, T. Simanouti, S. Nagakura, K. Kuratani, M. Tsuboi, H. Baba, and O. Fujioka, J. Am. Chem. Soc. **72**, 3490 (1950).
[11] R. M. Badger and H. Rubalcava, Proc. Natl. Acad. Sci. U. S. **40**, 12 (1954).
[12] W. T. Astbury, Nature **163**, 722 (1949).
[13] R. E. Marsh, R. B. Corey, and L. Pauling, Biochim. et Biophys. Acta **16**, 1 (1955).
[14] C. H. Bamford, L. Brown, E. M. Cant, A. Elliott, W. E. Hanby, and B. R. Malcolm, Nature **176**, 396 (1955).
[15] F. H. C. Crick and A. Rich, Nature **176**, 780 (1955).
[16] F. H. C. Crick and A. Rich in *Recent Advances in Gelatine and Glue Research* (Pergamon Press, London, 1957), p. 20.
[17] A. Rich and F. H. C. Crick, Nature **176**, 915 (1955).
[18] P. M. Cowan, S. McGavin, and A. C. T. North, Nature **176**, 1062 (1955).
[19] G. N. Ramachandran, Nature **177**, 710 (1956).
[20] L. Pauling, R. B. Corey, and H. R. Branson, Proc. Natl. Acad. Sci. U. S. **37**, 205 (1951).
[21] L. Pauling and R. B. Corey, Proc. Natl. Acad. Sci. U. S. **37**, 241 (1951).
[22] C. H. Bamford, L. Brown, A. Elliott, W. E. Hanby, and I. F. Trotter, Proc. Roy. Soc. (London) **B141**, 49 (1953).
[23] M. F. Perutz, Nature **167**, 1053 (1951).
[24] W. T. Astbury, Proc. Roy. Soc. (London) **B141**, 1 (1953).
[25] F. H. C. Crick, Nature **170**, 882 (1952).
[26] L. Pauling and R. B. Corey, Nature **171**, 59 (1953).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

6

Physicochemical Characterization of Macromolecules

PAUL DOTY

Department of Chemistry, Harvard University, Cambridge 38, Massachusetts

ALTHOUGH the x-ray diffraction method appears to be the only means capable of yielding the complete three-dimensional structure of large macromolecules (p. 94), an investigation of macromolecular structures in solution is important for two reasons. In a number of cases, the macromolecule is either not crystallizable or insufficiently crystallizable; consequently, x-ray diffraction methods cannot be applied. In the other cases, where x-ray structure determination is possible, physicochemical examination is necessary to show whether or not the configuration characteristic of the solid state remains intact in solution.

It is essential, therefore, that one find out from the physical methods as much as possible about these molecules in solution. These methods are complementary to the x-ray investigation which goes on in the solid state. In addition, they furnish unique information such as molecular weight and molecular-weight distribution. In general, the more that one can do to bridge these two approaches, the better.

Nature seems to deal primarily with polymeric structures in the production of macromolecules. Fortunately, there is not the chaos that might be present should all possible covalent bonds be joined together to make an infinite variety of large macromolecules. There has been, in fact, a great sorting out so that only three basic polymeric chains have been selected for wide use in living systems. Evolved over many millions of years are the polysaccharides, the polypeptide, and the nucleic-acid chains. These sometimes are tied together by cross-links, but this generally does not mar the basic simplicity of the long-chain structures which may, or may not, be folded in a particular way. It is these polymer chains, subject to many specific perturbations, that we wish to examine in the greatest detail that is practical.

SPECIAL PROBLEMS IN CHARACTERIZING MACROMOLECULES

One must consider at the outset three special and constantly recurring problems which arise in the characterization of macromolecules and have no simple counterparts in the field of small molecules. First to be mentioned is the fact that there may be a distribution of molecular weight. This does not usually occur in globular proteins, but it does in most other biological macromolecules. Consequently, the molecular weight is no longer a unique and simple number. One must deal instead with chains of varying lengths, but which are otherwise homogeneous. Each physical method that can be used for molecular-weight determination reflects a particular average of the molecular-weight distribution and as a result this average must be specified.

There are two widely used averages. The first of these is the number-average molecular weight—a very democratic form of averaging wherein each molecule is counted with a weight of one regardless of the actual weight. The second, the weight-average, corresponds not so much to democracy as to a more primitive arrangement wherein the importance of the molecule influences its count in accordance with its actual physical weight. The ratio of one of these to the other is a rough measure of the breadth of the molecular-weight distribution. This quality of having a distribution of molecular weight is sometimes referred to as "polydispersity."

Also to be considered is the problem of the specification of molecular size and shape. There are a variety of situations which arise from the basic polymeric chains. At one extreme, there can be found no periodic internal structure to the macromolecule whatsoever. It will become a randomly coiled molecule which is a rather common but degenerate form, since it does not have any specific configuration that is generally thought to be required for the functioning of many biological macromolecules. At the opposite extreme, many species of the three types of biological macromolecules can form helices (see Rich, p. 50), most commonly made of one, two, or three molecular strands, in which atoms are grouped in a periodic, inflexible array. Now, it is clear that the specification of size and shape must be treated differently in these two cases. For the case of the random coil, one will wish to measure the volume of space occupied by the molecule or some effective radius; in the case of the compact helical structure, one obviously will measure the length of the molecule and its mass, and this ratio of length to mass will characterize the particular helix. Therefore, the kind of parameter used to characterize the size of a macromolecule depends very much upon the form the molecule presents.

The final feature, which arises here and which has no counterpart with small molecules, is a space-filling concept in solution. If one had to deal only with dilute solutions of small compact spheres, then these particles could be considered as spaced at random with no overlapping and only rarely making contact. However, if these same molecules were not compact spheres, but random coils of the same mass, they would be expanded to such an extent that they would be continuously overlapping. This feature is maximal in the case of the

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

random-coil form, but it is significant for rigid, asymmetric macromolecules as well. The consequence of this is that, if physical measurements are to be used to characterize the individual properties of these molecules, the system must be diluted down to the point where molecules are clearly separated one from another. This requires a very high dilution which, in turn, puts a very severe requirement upon the sensitivity of the physical method to be employed. A number of physical methods which otherwise might be very attractive fail on this account. They are not sufficiently sensitive to respond to the very high dilutions necessary to allow the individual expression of the intrinsic properties of the single macromolecule.

THERMODYNAMIC METHODS

Keeping in mind these three special problems, a survey is undertaken now of what appear to be the most practical and general methods for obtaining quantitative information about the anatomical character of biological macromolecules. It is of interest to note that all of these methods have their origins in the work of eminent physicists at the turn of the century. The contributions of Einstein and Boltzmann appear in a number of places and our indebtedness to them is very great.

Osmotic Pressure

The thermodynamic methods, three in number, are considered first. Perhaps the widest known is that of osmotic pressures. It is one property of solution which, as it happens, gives a macroscopic response to the dilute macromolecular solutions that is sufficiently large to be practical. The osmotic pressure is a measure of the number of molecules of solute per unit volume in solution. Therefore, if the weight of the material per unit volume in solution is known, and if one measures the number by means of osmotic pressure, a number-average molecular weight can be obtained.

This was perceived in the latter part of the last century, but only recently has the exact and rigorous relation of the concentration-dependent behavior of osmotic pressure to that of gas pressure been demonstrated. This is important, because, in each one of these methods, particular attention must be paid to the way in which the physical property depends upon concentration, and so, in each case, measurements must be carried out on a series of dilute solutions and extrapolated to an infinitely dilute solution in order to isolate the property of the individual molecule.

In all of these thermodynamic methods, the concentration dependence bears a close analogy to that of the pressure in a gas. Thus, the pressure-volume relation of an imperfect gas has the following form

$$PV/nRT=1+B(n/V)+C(n/V)^2+\cdots, \qquad (1)$$

where B and C are the second and third virial coefficients. The analogy with the concentration dependence of osmotic pressure is exact. Osmotic pressure, π, replaces gas pressure, P, and concentration (mass per unit volume) divided by molecular weight, c/M, replaces number of moles per unit volume, n/V. In this way,[1] one obtains the relation

$$\frac{\pi}{cRT}=\frac{1}{M}+\frac{B}{M^2}c+\frac{C}{M^3}c^2+\cdots. \qquad (2)$$

At sufficient dilution, the terms in c^2 and higher will be negligible. Under these conditions, a plot of π/cRT against c yields a straight line whose intercept is the reciprocal of the number of average molecular weight and whose slope is equal to B/M^2.

In small molecules of gases, the virial coefficient often is thought of in terms of excluded volume—that is, the amount of space not available for the center of one molecule because of the volume of another molecule. In the domain of macromolecular solutions, this effect can become very much larger because of the extra-large volume-filling capacity of some forms of macromolecules. However, if the molecules have an attraction for each other, this over-all effect becomes smaller and even can become negative immediately before phase separation. On the other hand, if there is a large electric repulsion, the over-all effect can become very, very large making it impossible to establish the curve with precision.

The range in which the osmotic-pressure method can be applied is from molecular weights of about 10 000 upward to a value which depends upon the precision desired—usually, a few hundred thousand. The lower limit is set by membrane permeability.

As an illustration, some measurements on collagen solutions[2] are shown in Fig. 1. The molecular weight obtained from the reciprocal of the intercept at the concentration axis is 310 000, and the scatter shows the uncertainty is of the order of 15%. This is about the upper limit for practical use of this method in aqueous solutions. The second virial coefficient can be evaluated from the slope, and, in this case, is found to arise almost entirely from the excluded volume effects.

Light Scattering

Light scattering, the second thermodynamic method, is considered next. Concentrating only upon its thermo-

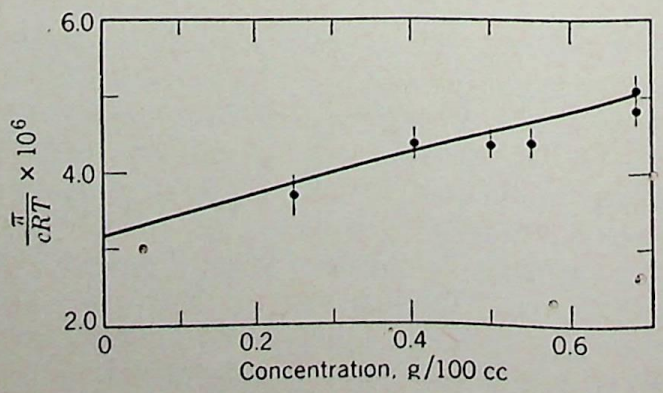


FIG. 1. Osmotic pressure of collagen solutions at 2°C.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

dynamic aspect, compounded, of course, by the central electrostatic features which go along with dipole scattering, one finds again a close relation to the solution of the corresponding problem in gases.

Rayleigh's analysis of the problem of light scattering from dilute gases, worked out nearly a century ago, took the following form. The removal of light from an incident beam by scattering may be expressed quantitatively by means of the turbidity, τ, defined as

$$I = I_0 e^{-\tau l}, \tag{3}$$

where I_0 and I are the intensities of a parallel beam of light before and after passing through a length, l, of the gas. Alternatively, the scattering can be characterized by measuring its intensity at any angle θ, i_θ, and a distance, r. When θ is 90°, τ and i_{90} are related in the following manner:

$$\tau = \frac{16\pi}{3}\frac{i_{90}r^2}{I_0} = \frac{16\pi}{3}R_{90}, \tag{4}$$

where R_{90} represents the reduced form of the scattering intensity and is known as the Rayleigh ratio. This ratio was shown to be for any angle

$$R_\theta = \frac{2\pi^2(n-1)^2}{\lambda^4}\frac{1}{\nu}(1+\cos^2\theta), \tag{5}$$

where n is the refractive index of the gas, λ_0 the wavelength of light in vacuum and ν the number of molecules per cc.

The relevance of this to scattering from a solution of macromolecules is due to Einstein[3] and Debye.[4] At very great dilution, the same results hold except that the refractive index of the solvent, n_0, replaces unity in Eq. (5). This then can be rearranged to give:

$$R_{90} = \frac{2\pi^2 n_0^2[(n-n_0)/c]^2}{N\lambda_0^4}cM = KcM. \tag{6}$$

From this it can be seen that the scattering as measured by R_{90} is proportional to the molecular weight (the weight average) and the proportionality factor is experimentally determinable.

This result holds for a collection of independent scattering centers. In real macromolecular solutions, the space-filling property, previously referred to, produces a concentration-dependent correlation between scattering centers. Einstein showed this could be taken into account in a manner that now can be expressed as follows:

$$K\frac{c}{R_{90}} = \frac{1}{M} + 2\frac{B}{M^2}c. \tag{7}$$

One sees here a complete analogy with osmotic pressure except for a factor of 2 and the type of molecular-weight average that is obtained. Again, scattering measurements must be made at a number of concentrations and upon extrapolation yield the desired intercept.

In contrast to osmotic pressure, the sensitivity of this method increases with molecular weight. However, when the molecular size becomes significant with respect to the wavelength of the light, the angular dependence of the scattered light is affected in a manner discussed in the next section, and this must be taken into account. Finally, it should be noted that, since osmotic pressure and light scattering provide different molecular-weight averages, their combined results provide a measure of the molecular-weight distribution.

Sedimentation Equilibrium

The third and final thermodynamic method is the various exploitations of sedimentation equilibrium that are found to be useful in the study of macromolecules. It is very likely that this will turn out to be the most versatile of all the tools which one can have for the physical investigation of macromolecules. The key to these methods is found in the high speed of the ultracentrifuge, which allows one to observe optically the distribution of macromolecules in a cell in which the gravitational field is increased to the order of 100 000 times that of gravity. The versatility of the modern ultracentrifuge allows one to carry out a number of quite different kinds of measurements. In one case, one may allow the field to be exerted on the molecules in the solution until they have redistributed themselves at equilibrium.[5] This would then correspond with the distribution of gas molecules in the earth's atmosphere, and the analysis of the situation can be readily adapted to yield an average molecular weight of the macromolecules distributed in the much greater field provided by the ultracentrifuge.

The basis of determining the molecular weight in this way can be seen if one recalls the relation between the pressure p and the height h in the earth's gravitational field g. That is,

$$\frac{dp}{p} = -\frac{Mg}{RT}dh. \tag{8}$$

In the ultracentrifuge cell, the pressure is replaced by the concentration, the molecular mass M is corrected for the buoyancy owing to the solvent by multiplying by $(1-\bar{v}p)$, where $\bar{v}p$ is the ratio of solvent to solute density, h is replaced by x, the distance from the center of rotation, and g is replaced by the centrifugal force w^2x. These exchanges lead to

$$\frac{dc}{c} = \frac{M(1-\bar{v}p)w^2x\,dx}{RT}. \tag{9}$$

If there is only one macromolecular species present, its molecular weight can be determined if the concentration is known at only two places in the cell. When a distri-

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

bution of molecular weights exists, its weight-average molecular weight can be obtained most readily.[6] Again, the observed molecular weight is found to be dependent upon the concentration in a manner governed by the second virial coefficient.[6] Consequently, the reciprocal of the apparent value must be plotted against concentration and the true value of M_w determined from the reciprocal of the intercept at zero concentration.

Now, in actual practice this method is not widely used because a number of days are required to attain the equilibrium distribution of the macromolecules in the solution in most cases. However, the smaller the molecular weight of the macromolecule, the sooner such distribution is attained; consequently, it is practical for molecules whose molecular weights are only a few thousand. Waugh[7] has pioneered in devising a particular kind of ultracentrifuge cell containing a retracting partition, which greatly facilitates this kind of application. By determining the concentration of the two parts, above and below the partition, one can obtain the molecular weight.

Within the last three years, an alternative version of the sedimentation-equilibrium method has come into use that promises to have wide utility. It is the Archibald approach to equilibrium method.[8] Kegeles[9,10] has been chiefly responsible for recognizing and developing its potential. Its basis is both novel and simple. The condition on which the general method of sedimentation equilibrium is based is that at equilibrium the net flux of solute species across any plane within the solution and perpendicular to the radius is zero. Archibald pointed out about 10 years ago that this condition of no net transport through a boundary perpendicular to the direction of sedimentation was always valid at the bottom and top of the solution, that is, at the meniscus and the cell bottom. Consequently, very soon after speed is attained, one can determine the initial redistribution of solute near the meniscus, or near the bottom of the cell; and from this one can derive the weight-average molecular weight by use of Eq. (9), together with appropriate extrapolation. Operationally, one now has precisely the opposite situation, a very short run instead of a very long run.

This method appears to be applicable over a wide range of molecular weight, and fortunately it does well in the region of 1000 to 10 000 where most other methods fail. As one illustration of this, I have chosen some work that we have done on determining the molecular weight of polypeptides and correlating this with intrinsic viscosity.

Before the Archibald method was available, it had been possible to show that the weight-average molecular weights of poly-γ-benzyl-L-glutamate were related to the intrinsic viscosity in the expected manner. That is, on a double logarithmic plot, a linear relation was found as indicated by the open circles in Fig. 2. We were unable to carry measurements below 20,000, however, and extrapolation was precarious. When the Archibald results were obtained (filled circles), the proper extension down to molecular weights of 1000 could be made as shown.[11]

INTERFERENCE OF SCATTERED LIGHT AND MOLECULAR SIZE

If the macromolecules are quite small as compared to the wavelength of light, and if the incident light is vertically polarized, the scattering will be the same at all angles; that is, $R_\theta = R_0$. However, for larger molecules with dimensions exceeding 200 or 300 A, interference arises from light scattered from different parts of the same particle. As a consequence, the scattering is diminished, the effect increasing with the scattering angle θ. The effect vanishes at zero angle and, as a consequence, measurements over the accessible angular range (usually 30 to 135°C) can be extrapolated to give R_0 which is required for molecular-weight determination. The character of the angular dependence, however, is of particular interest because it reflects the distribution of matter within the scattering article. The angular dependence can be separated from the concentration dependence because the following relation is generally valid

$$\frac{Kc}{R_\theta} = \frac{1}{MP(\theta)} + 2Bc. \qquad (10)$$

That is, the angular dependence can be shown to enter only as a function, $P(\theta)$, known as the particle-scattering factor equivalent to the square of the structure factor in x-ray diffraction. The experimental determination of $P(\theta)$ is used to obtain information of dimensions and shape through the relation

$$P(\theta) - 1 - \frac{\rho^2}{3}\left[\frac{4\pi \sin(\theta/2)}{\lambda'}\right]^2 + \cdots, \qquad (11)$$

where λ' is the wavelength of light in solution and ρ is

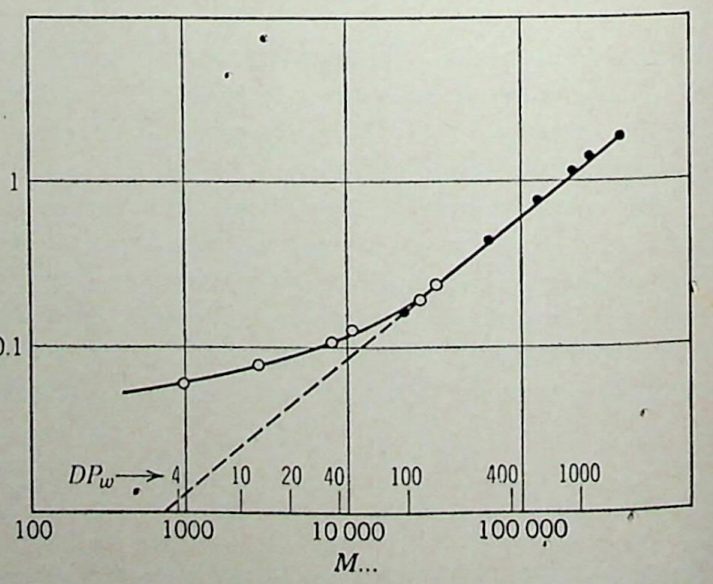


FIG. 2. A double logarthmic plot of intrinsic viscosity measured in dichloroacetic acid against weight-average molecular weight of poly-γ-benzyl-L-glutamate.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

radius of gyration of the macromolecule. This radius of gyration is related to the dimensions of simply shaped particles. For example, the length of a rod-like molecule is $(12)^{\frac{1}{2}}\rho$. The root-mean-square end-to-end distance $(\langle r^2\rangle)^{\frac{1}{2}}$ of a randomly coiled polymer is given by $(6)^{\frac{1}{2}}\rho$. For quite large macromolecules, higher terms in Eq. (11) become significant in such a way that shape, as well as the size, can be independently determined or at least estimated.

As an illustration of these methods, one can consider a particular fraction of cellulose nitrate.[12] Osmotic-pressure measurements plotted according to Eq. (1) yielded a value of 234 000 for M_n. Light-scattering measurements showed substantial angular dependence. A double extrapolation against concentration and $\sin^2(\theta/2)$ in accordance with Eqs. (10) and (11), known as a Zimm plot, showed that $M_w = 400\,000$ and $(\langle r^2\rangle)^{\frac{1}{2}} = 1500$ A (Fig. 3). This indicates a rather broad molecular-weight distribution for a sample that has already been fractionated and the size is indicative of a quite extended random coil. This is evident when the following is considered. Each monomer unit has a molecular weight of 294 and a length of 5.15 A. Hence, the number of units (degree of polymerization) making up a chain of 400 000 molecular weight is 1350. Completely extended, this would be 6950 A in length. By comparing this with $(\langle r^2\rangle)^{\frac{1}{2}} = 1500$ A, the extent of coiling is readily visualized. If there were no hindrances to rotation at the glycosidic linkages, the value of $(\langle r^2\rangle)^{\frac{1}{2}}$ would be several times smaller. The extent to which this molecule is extended because of steric and potential hindrances to rotation is the highest yet found for single chains. One can appreciate that this "stiffness" is put to good use in the biological role of this material since this will not only give rigidity to the crystallites of cellulose, but will make firm the amorphous regions in between the crystallites. Moreover, this natural tendency toward rod-like behavior in localized regions of the chain is effective in lowering the entropy of melting, and thereby contributes to the high melting point and hence the negligible solubility of cellulose.

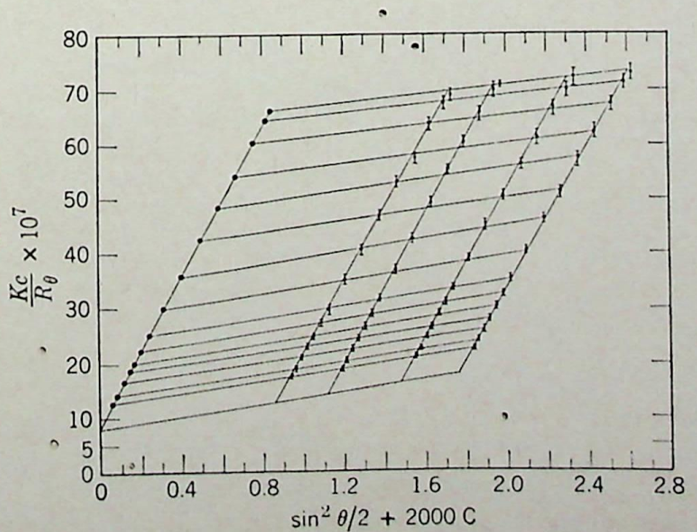


FIG. 3. Cellulose-nitrate fraction Ab in acetone at 25°C.

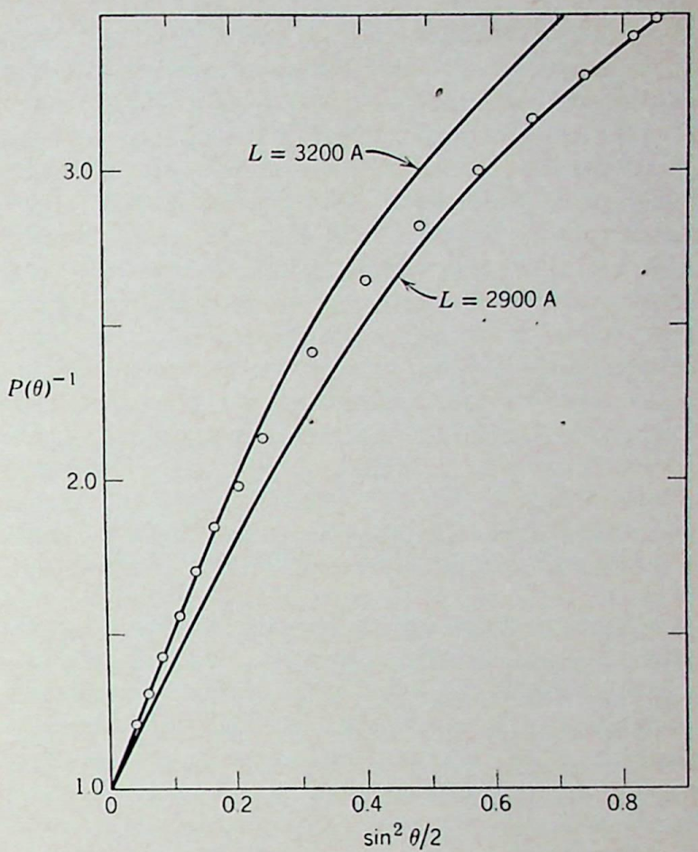


FIG. 4. Reciprocal particle-scattering factor of tobacco mosaic virus solution: sample *A*-4, experimental points,—theoretical scattering curves.

The Zimm plot for tobacco mosaic virus is shown in Fig. 4. It can be seen that the experimental points are bounded by the theoretical scattering curves for rods (having negligible diameter) of 2900 and 3200 A. Boedtker and Simmons[13] were able to conclude from these data that the length was 3000 ± 50 A. In contrast to the complete linearity of the plot for the randomly coiled cellulose derivative in the previous figure, one finds here a pronounced downward curvature characteristic of rod-shaped scattering elements. From the molecular weight of 39.5 million obtained, from the intercept and the length and density, an effective diameter of 150 A can be computed. Thus, the shape and dimensions of this virus are completely determined from light-scattering measurements.

An application of light-scattering studies to collagen[2] is shown in Figs. 5 and 6. In the former, the Zimm plot for the native-collagen molecule in solution is shown. The downward curvature indicates a rod-like shape and the quantitative interpretation of the results show that the molecule is 3000 A long and has a molecular weight of 360 000. The diameter in this case is only 13.5 A and, as described by Rich (p. 50), consists of only three polypeptide chains in a helical arrangement. As in other cases which are dealt with in the paper by Rich, this macromolecule is essentially a one-dimensional crystallite. Hence, it should melt on raising the temperature.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

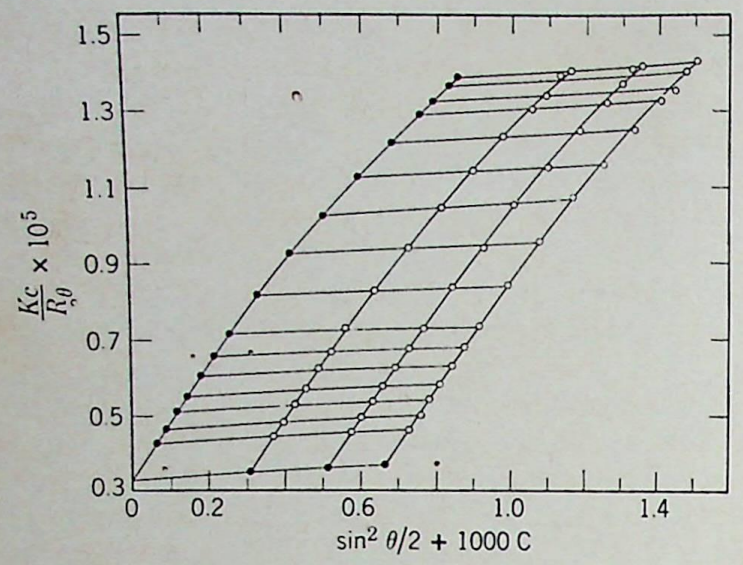


FIG. 5. Collagen (ichthyocol) in citrate buffer (*p*H 37) at 15°C. M_w=310 000; P=955 A; L=3300 A; M/L=94.

Upon melting, the secondary bonding holding the three chains together disappears and the "melted" or denatured state should consist of the three polypeptide chains in randomly coiled configurations. This configurational transition can be observed directly in light scattering. In Fig. 6, the reciprocal envelopes are plotted before and after heating. The intercept is seen to be nearly three times higher indicating a correspondingly lower molecular weight, and the slope is likewise greatly diminished indicating a much smaller spatial extent. The curvature is also gone, consistent with the assumption of a randomly coiled configuration.

Light-scattering studies of deoxyribonucleic acid ((DNA) have been particularly useful in showing the nature of these molecules in solution. These macromolecules are so highly extended in space that very great dilution is required in their physical characterization, and as a consequence few methods could be properly applied. A typical light-scattering result is shown in Fig. 7. The slope relative to the intercept is seen to be much greater than those encountered before. For this case, its interpretation as a coiled molecule leads to an estimate of 7000 A for the average end-to-end length and a molecular weight of 8 000 000. The contour length of this molecule is about 40 000 A on the basis of the Watson-Crick model. Thus, it is only very modestly curved and the mild downward curvature reflects its intermediate status between a coiled and rod-like shape. Actually, the size of this molecule is greater than present light-scattering techniques can handle because the extrapolation should be based on measurements down to 5°C in order to eliminate the effects of polydispersity on the final answer. Thus far, however, the errors have not been serious because the polydispersity has been such as to justify a linear extrapolation from the higher angles, but this good fortune cannot be expected to hold in general.

The use of the angular dependence of scattered light in this manner requires only the ratio of particle size to wavelength to be such that the interference is first order. This situation holds in the region of x-rays as well, and as a consequence the scattering of long x-rays can lead to the assignment of dimensions of the order of hundreds of A, just as light scattering deals with dimensions of a few thousands A. Time does not permit an examination of this, but reference should be made to the work of Beeman[14] and Luzatti[15] in this connection.

HYDRODYNAMIC METHODS

The motions of a macromolecule in solution can be resolved into those of translation and rotation. If the motion is simply that of diffusion, one terms it translational or rotatory diffusion and characterizes it by a diffusion constant D for translation and Θ for rotation. If the motion arises from an imposed gradient, two other situations occur. If the imposed force is that of a gravitational gradient as in the ultracentrifuge, sedimentation is observed and one characterizes it by a sedimentation constant s. If the imposed force arises from a hydrodynamic gradient, the molecule is caused to rotate with a definite bias instead of at random. The dissipation of energy that this produces gives rise to an increase in viscosity over that of the solvent. The fractional increase is known as the specific viscosity, and, when divided by concentration and extrapolated to zero concentration, it becomes known as the intrinsic viscosity $[\eta]$.

These four types of motion are all the result of a certain amount of resistance to an applied force (or couple). This resistance is hydrodynamic in nature and arises from the size and shape of the molecule. On very general grounds, it can be argued that for a given molecule the resistance can be characterized by a frictional factor f, which is the same in all four cases. Consequently, two developments are possible. Theoretical

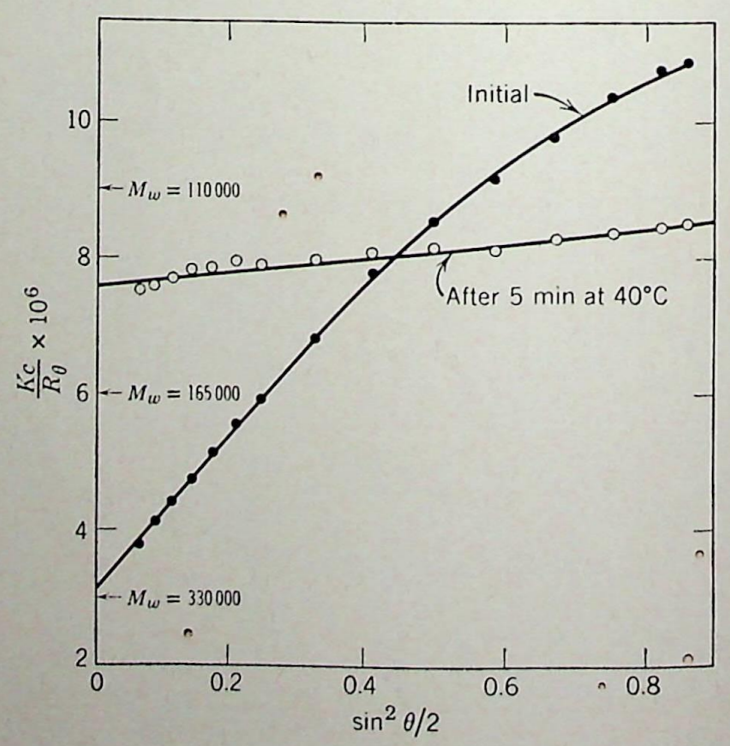


FIG. 6. Denaturation of collagen in solutions. ● = before denaturation; ○ = after denaturation.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

investigations can aim at calculating the frictional factor for various macromolecular models as a function of dimensions and provide, in this way, a basis of obtaining dimensions from measuring one or more of the foregoing quantities. On the other hand, the measurement of any two quantities provides at least the possibility of eliminating f and obtaining the molecular weight. This is illustrated in the well-known Svedberg equation. Here, the expressions for D and s are combined to eliminate f as shown.

$$D=kT/f, \qquad s=M(1-\bar{v}\rho)Nf, \tag{12}$$

$$M=\frac{sRT}{D(1-\bar{v}\rho)}. \tag{13}$$

Thus, if the partial specific volume $\bar{v}$ and the solvent density ρ are determined, the molecular weight can be obtained from measurements of s and D.

The difficult and time-consuming nature of the measurement of D, particularly for chain-like and rod-like macromolecules, has led to a search for other means of achieving the same result by using the much more easily measured quantity, the intrinsic viscosity $[\eta]$, in its place. This can be done rigorously for ellipsoids, making use of the work of Perrin and Simha. The result is

$$M=\left[\frac{s[\eta]^{\frac{1}{3}}\eta_0 N}{\beta(1-\bar{v}\rho)}\right]^{\frac{3}{2}}, \tag{14}$$

where the constant β is a slowly varying quantity dependent only upon the axial ratio. Its values range from 2.12 for spheres to 3.50 for infinite axial ratio. The axial ratio can always be determined from $[\eta]$ with enough precision to permit the correct choice of β.

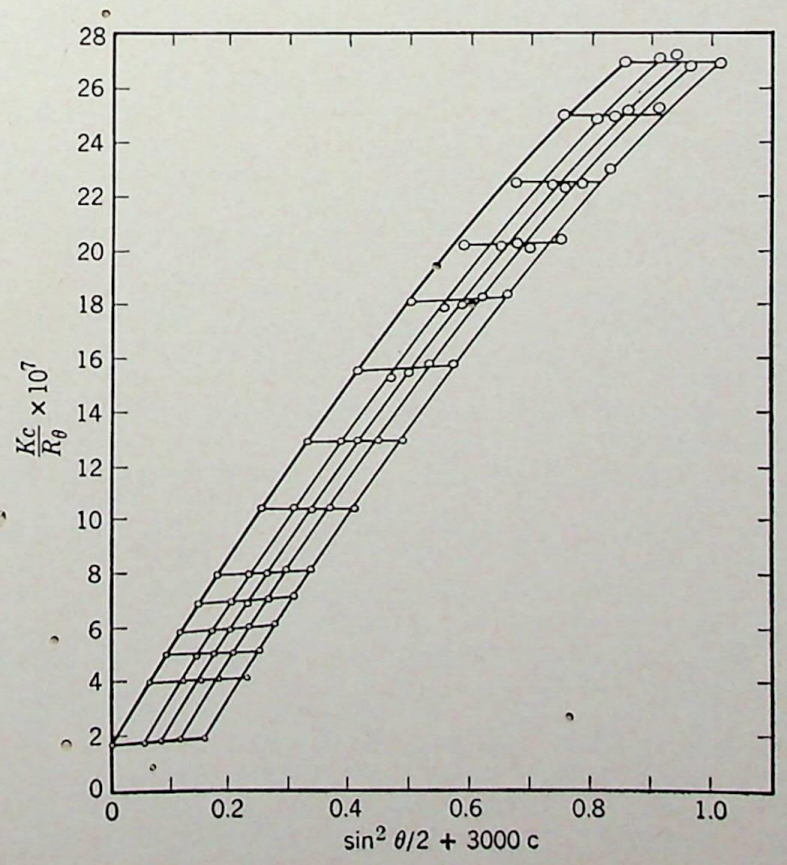


FIG. 7. Light scattering of DNA.

The usefulness of intrinsic-viscosity determinations is by no means limited to Eq. (14). Quite the contrary. Its application stems from a series of impressive theoretical derivations that have been made relating it to size and shape. The first of these was owing to Einstein who showed that the intrinsic viscosity of compact (unpenetrated by solvent) spheres in solution was 2.5 times their specific volume in units of cc/g. This value represents a lower limit: almost all values observed for macromolecules are much larger. For example, 4 to 10 for a number of globular proteins, 10 to 100 for a number of proteins known to be asymmetric, and extremely high values in the case of collagen (1200) and deoxyribonucleic acid (7000). Later theories showed that these higher values could all be accounted for in terms of two effects: asymmetry of the macromolecule, or solvent immobilization. The latter effect could arise from hydration owing to the binding of water (or other solvent) at specific sites on the macromolecule or by loosely incorporating it with the domain of the macromolecule through swelling. Thus, a typical randomly coiled polymer pervades a domain that is perhaps 100 times or more greater than it would occupy in a compact form; yet, the solvent in this domain moves with the polymer and is, therefore, hydrodynamically immobilized. As a consequence, the intrinsic viscosity measures this volume of immobilized material, and, in a rough way, its numerical value expresses in units of cc/g the ratio of the volume of the domain to the mass of the polymer molecule. Extremely useful developments in the last few years have led to the recognization of the relation between intrinsic viscosity (and sedimentation constant as well) and the mean end-to-end length, $(\langle r^2\rangle)^{\frac{1}{2}}$, of polymer chains.[16] As a consequence, the size of such polymer molecules can be determined from the intrinsic viscosity, if their molecular weight is known. Moreover, the large variations in intrinsic viscosity observed when the solvent or temperature is changed are seen to be the result of swelling and shrinking of the polymer coils. Finally, the previously observed empirical relation between intrinsic viscosity and molecular weight for a series of homologous polymers,

$$[\eta]=KM^a, \tag{15}$$

becomes understandable from this point of view. The exponent in this relation is a constant with a value between 0.5 and 1.0 for any given polymer-solvent system. These limiting values correspond to tightly coiled molecules with many intramolecular contacts (0.5) and to very highly swollen molecules (1.0), with most actual cases distributed in between.

In the case of rigid particles, the exaltation of the intrinsic viscosity arises from the increased energy dissipation resulting from the end-over-end rotations in solution. The analysis of this problem, particularly by Simha,[17] showed that the intrinsic viscosity was a func-

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

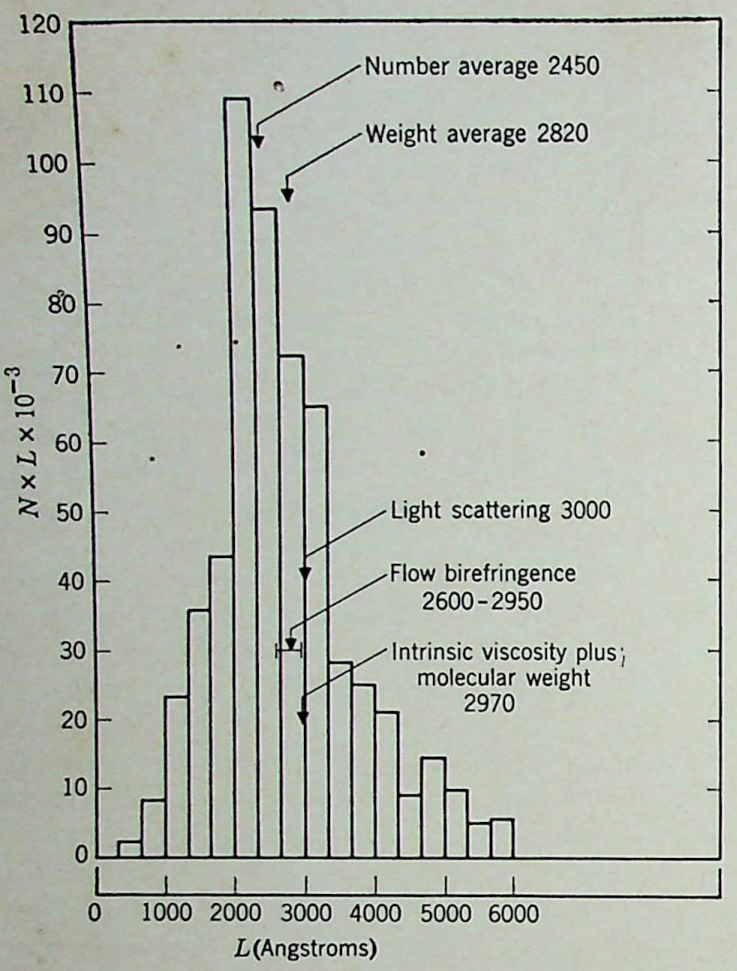


FIG. 8. Weight distribution of lengths of ichthyocol macromolecules as measured from electron micrographs compared with results from other methods. Total number represented is 238 [from C. E. Hall and P. Doty, J. Am. Chem. Soc. **80**, 1269 (1958)].

tion only of the axial ratio of the particle and that this dependence rapidly approached that of the square of the axial ratio. In many typical proteins, one is faced with the possibility that the intrinsic viscosity reflects both molecular asymmetry and moderate amounts of hydration. The resolution of this problem has been worked out by Oncley.[18]

The final point to make on intrinsic viscosity is about its ease of measurement relative to the other three hydrodynamic properties.

The determination of the rotatory-diffusion constant from streaming birefringence of flow is a rather specialized technique and is limited to rigid particles. In these cases, provided the length of the molecules exceeds a few hundred angstroms, the length and often the length distribution can be measured with considerable accuracy.

As a single illustration of the methods discussed in this section, as well as a demonstration of their self-consistency, the results on soluble collagen are presented in Table I.[2] The intrinsic viscosity was 1150 cc/g, the sedimentation constant was 2.96 Svedbergs and the distribution of rotatory-diffusion constants corresponded to a range in length from 2500 to 2950 A.

This brief discussion of the physical determination of the characteristics of macromolecules would not be complete without the mention of a new technique that has, within the last two years, become practical for asymmetric, rigid macromolecules. This is the perfection of the direct viewing of such macromolecules in the electron microscope by Hall.[18] Thus, nucleic acids, some proteins, and even simple polypeptides in the α-helical configuration have been measured and found to be in agreement with the results of other physical methods applied in solution.[19] As one illustration of this, shown in Fig. 8 is the weight distribution of the lengths of collagen molecules compared with the average values shown in Table I. The agreement is seen to be quite good. It is possible that the distribution was somewhat broadened as a result of some damage in spraying and shadowing the electron-microscope preparation, but the effect has been modest. F. O. Schmitt deals with the way in which these macromolecules are united in collagen fibrils and the extent to which the dimensions found here are compatible with his studies (p. 349).

TABLE I.

Method	Mol. Wt.	Length A	Diameter A
Osmotic pressure (M_n)	310 000	...	...
Light scattering (M_w)	345 000	3100	13.0
Intrinsic viscosity and M_w	...	2970	13.6
Sedimentation and viscosity	300 000	...	12.8
Flow birefringence and viscosity	350 000	2900	13.5

BIBLIOGRAPHY

[1] The rigorous derivation of Eq. (2) may be found in the work of W. G. MacMillan, Jr., and J. E. Mayer, J. Chem. Phys. **13**, 276 (1945).
[2] H. Boedtker and P. Doty, J. Am. Chem. Soc. **78**, 4267 (1956).
[3] A. Einstein, Ann. Physik **33**, 1275 (1910).
[4] P. Debye, J. Phys. & Colloid Chem. **51**, 18 (1947).
[5] T. Svedberg and K. Pedersen, *The Ultracentrifuge* (The Clarendon Press, Oxford, 1940).
[6] M. Wales, J. W. Williams, J. O. Thompson, and R. H. Ewart, J. Phys. & Colloid Chem. **52**, 983 (1948).
[7] D. A. Yphantis and D. F. Waugh, J. Phys. Chem. **60**, 630 (1956).
[8] W. J. Archibald, J. Phys. Chem. **51**, 1204 (1947).
[9] S. M. Klainer and G. Kegeles, J. Phys. Chem. **59**, 952 (1955).
[10] S. M. Klainer and G. Kegeles, Arch. Biochem. Biophys. **63**, 247 (1956).
[11] J. C. Mitchell, A. E. Woodward, and P. Doty, J. Am. Chem. Soc. **79**, 3955 (1957).
[12] A. M. Holtzer, H. Benoit, and P. Doty, J. Phys. Chem. **58**, 624 (1954).
[13] H. Boedtker and N. S. Simmons, J. Am. Chem. Soc. **80**, 2550 (1958).
[14] H. N. Ritland, P. Kaesberg, and W. W. Beeman, J. Chem. Phys. **18**, 1237 (1950).
[15] M. Champagne, V. Luzzati, and A. Nicolaieff, J. Am. Chem. Soc. **80**, 1002 (1958).
[16] P. J. Flory, *Principles of Polymer Chemistry* (Cornell University Press, Ithaca, New York, 1953), p. 595.
[17] R. Simha, J. Phys. Chem. 44, 25 (1940).
[18] C. Hall, J. Biophys. Biochem. Cytol. **2**, 625 (1956).
[19] C. Hall and P. Doty, J. Am. Chem. Soc. **80**, 1269 (1958).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

REVIEWS OF MODERN PHYSICS VOLUME 31, NUMBER 1 JANUARY, 1959

7
Polyelectrolytes

STUART A. RICE

Department of Chemistry and Institute for the Study of Metals, The University of Chicago, Chicago 37, Illinois

I. INTRODUCTION

THE past decade has witnessed an extensive development in two directions of the theory of electrolyte solutions. On the one hand, the investigations of Mayer[1] and of Kirkwood and Poirier[2] finally have established from first principles the validity of the Debye-Huckel limiting law. These studies provide a formal framework for extending the theory of electrolytes to concentrations higher than the traditional "slightly polluted water," and they provide insight into the nature of the approximations which make valid the use of the Poisson-Boltzmann equation in dilute solution. The recent works of Onsager and Kim,[3] of Fuoss and Onsager,[4] of Falkenhagen and Kelbg,[5] and of Pitts[6] have extended, within the framework of the Poisson-Boltzmann equation, the theory of transport processes to include the effects of finite ion size. While it is doubtful if the validity of the Poisson-Boltzmann equation permits further development, the inclusion of finite ion size does permit the description of dissipative processes, in terms of an ion-size parameter, to be extended to much higher concentrations than heretofore, and it suggests a more satisfactory interpretation of the varying behavior of different strong electrolytes. Further extension of the theory of dissipative processes probably requires the more powerful tools of the Mayer theory combined with the general statistical theory of transport.[7]

On the other hand, there has been an almost explosive development of the theory of very highly charged macromolecular ions. Whereas, prior to 1940, there existed almost no quantitative theory of colloid solutions, and whereas the measurements of Kern represented the only investigations of flexible polyelectrolytes,[8] there exists now a number of theories for the equilibrium properties of these substances, and there have been several attempts at the theory of transport. In all instances, the theory of polyelectrolytes has been developed within the framework of the Poisson-Boltzmann equation; in some cases, the analysis uses a linearized form, while in others the analysis adopts different approximations equivalent to the assumption that the domain occupied by the polyion has zero net charge. Although there are many difficulties associated with the current theory, and, though qualitative or semiquantitative agreement with experiment is all that can be attained at present, it may be asserted safely that the broad features of the basic physical processes responsible for the observed phenomena have been delineated and that the major problems are quantitative rather than qualitative. In this article, the theory of polyelectrolyte solutions is reviewed briefly, with emphasis on the basic physical processes. The available experimental data suitable for comparison with the theory, and the directions which future developments are likely to take are mentioned only cursorily, owing to limitations of space.

II. GENERAL SURVEY AND QUALITATIVE PICTURE

The fundamental theory of rigid, impenetrable macroions first was published in 1948 by Verwey and Overbeek[9] in their well-known monograph; it has been refined continuously by Levine,[10–12] by Booth,[13] and by Kirkwood and co-workers.[14,15] Briefly, the model adopted for the interaction between particles represents the macroion as a rigid sphere with a specified surface-charge density. The high surface charge tends to force the counterions to cluster close to the surface of the macroion, creating a thin double layer which shields the surface charges from similar charges on other colloidal particles. The net repulsive force between two macroions usually is obtained from a suitable solution of the Poisson-Boltzmann equation, although more sophisticated methods also have been used.[14–16] The fact that the macroion is rigid and impenetrable simplifies the problem to the extent that almost no features are intrinsically different from the low molecular weight, strong electrolyte case. There are, however, qualitative differences in the relative importance of the various contributions.

For colloidal ions with one type of charge, one may imagine the free energy of building the double layer to consist of four contributions. First, if the charge is due to adsorption or desorption of ions from the solution, there is an intrinsic chemical potential change due to the dissociation reaction. To this must be added the sum of the chemical potentials of all of the species in solution, the difference in self-energy (interaction of central ion and atmosphere) of an adsorbed ion from a free ion, and the electrical free energy of charging up and assembling the double layer. When considering the interaction of two macroions, it is important to note that the overlap of double layers causes a redistribution of small ions in the solution with the net result that the mutual interactions of the small ions are changed. Thus, in addition to the expected screened Coulomb force between macroions, there is a force arising from the change in interaction energy of the small ions as they redistribute under the influence of the overlapping double layers.

When consideration is given to the possibility that charges of both signs may exist on the surface of the

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

molecule, a new mechanism of interaction between macroions becomes apparent.[16] If there is a net difference in the numbers of positive and negative charges, at the isoelectric point there will be a number of uncharged groups otherwise identical in chemical character with some of the charged groups. For one of the types of charge, there exists under these conditions a large number of different possible arrangements of the charges on the surface of the molecule. A number of these arrangements is of approximately equal electrostatic energy and one may expect the charges to fluctuate in occupation of first one and then another set of sites. If two such molecules are brought close to each other, the charges on one ion tend to polarize the charges on the other in such a manner as to separate like charges. This is possible if there exist surface-charge distributions of roughly equal energy and the resultant decreased repulsion is, of course, equivalent to a net attraction, since the net charge of each particle is zero. It has been postulated that this fluctuation force is operative in enzyme reactions.[17]

Another fundamental difference between macroions and small ions is that, in the former case, consideration must be given to the mutual interactions of charges on the surface of the ion. The magnitude of these interactions depends upon the geometric arrangement of the charges, upon the dielectric constant, and upon the salt concentration, and readily may be ascertained experimentally from the titration curve. Early attempts to characterize this electrostatic energy made the crude approximation of smearing out the discrete-charge structure and applying the Debye-Huckel theory in substantially its original form. Refinements, such as allowing the ion to be partially penetrable by solvent, alter the model only slightly.[18] On the other hand, direct calculation[14,15] of the electrostatic energy for discrete-charge distributions reveals a marked dependence of the titration curve upon the details of the distribution of charges, as was to be expected. Physically, this corresponds to the fact that neighboring acid and base groups tend to become zwitterions, whereas neighboring groups of the same kind resist ionization because of the large repulsive interactions created by such ionizations. Consequently, the amount of work required to ionize a group depends markedly upon the group environment, and a smeared-out charge distribution gives an erroneous, even if fortuitously numerically accurate, picture of the molecule.

In the case of flexible polyelectrolytes, attention has been focussed almost entirely on the interactions between the charges of the polyion and on the relationship between configurational and thermodynamic properties and ionic strength, degree of neutralization, and so forth. Interactions between polyions only now are being investigated in detail.

A basic step in the study of linear polyelectrolytes was taken in 1948 by Kuhn, Künzle, and Katchalsky,[19] who described the configuration of a polyion as that of a chain which was randomly coiled with the restriction that its mean end-to-end distance be such as to minimize the sum of the configurational and electrostatic free energies. This simple model was able to predict qualitatively the size changes accompanying changes in charge on a polymer. This treatment was approximate, however, in that it assumed the charges on the polymer to have an interaction energy determined only by the end-to-end distance, rather than by all of the distances between charges on the chain. Moreover, it was not shown that a screened Coulomb potential was the proper expression to use in deriving the interaction energy between a pair of fixed charge in a solution containing mobile ions.

In the same year, Hermans and Overbeek[20] proposed a model in which the polymer was regarded as a porous spherical-charge distribution, whose radius was to be varied to minimize the total free energy. By using a linearized form of the Poisson-Boltzmann equation, they computed the electrostatic free energy without using the screened Coulomb potential. Although the Hermans-Overbeek model avoids explicit consideration of the chain configuration, it predicts the size changes of polyions moderately well. But neither of the models discussed thus far was conspicuously successful in interpreting the titration curves of polyelectrolytes.

Two years later, Huizenga, Grieger, and Wall[21] reported experiments which, for the first time, fully showed the importance of ion binding in polyelectrolyte solutions. They found that up to 60 percent of the sodium ions counter to a carboxylate polymer moved with it in electrophoresis. Since these polymers are thought to be approximately free draining under the conditions of this experiment, one concludes that the counterions are not merely within the volume of the polymer, but are relatively tightly bound to it.

The foregoing work, and contributions by Flory,[22] by Kimball, Cutler, and Samelson,[23] by Osawa, Imai, and Kagawa,[24] and by others, led to the proposal of a model[25] which was new in that (a) the interactions between charges of the polymer influenced its local configurational properties as well as its end-to-end distance, (b) these interactions affected the tendency toward ionization at each ionizable group, making the various ionizable groups of a polymer interdependent, rather than independent, as heretofore postulated, and (c) the interactions were large enough to cause important amounts of binding of counterions, even though negligible amounts of binding would occur if the ionizable groups did not interact. In implementing this model, it is necessary to allow for the interrelations among its several features. In particular, to determine the mean size, degree of ionization, and degree of binding of a polyion in a solution of given pH, counterion concentrations, and ionic strength, one would determine first the free energy as a function of all six of these

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

variables, and then minimize this with respect to the first three, which are not controllable externally. The expression for the free energy must reflect the fact that not all chain configurations of the same end-to-end distance will have the same electrostatic interaction, nor will all arrangements of the same total number of charges be equienergetic for a given chain configuration. A further discussion of these and related points is presented after a more detailed definition of the molecular model.

To specify the model more completely, consider a weak polyacid, each functional group of which can be un-ionized (zero charge), ionized (−1 charge), or ion paired (zero charge). It is assumed that an intrinsic "chemical" free energy change can be assigned to the conversion of an ionizable group, from one to another of these charge states, so that, ignoring interactions among the charges, the ionization and binding would be describable by equilibrium constants. The binding is thus treated phenomenologically. The ion pairs introduced here may be described as "site bound," so as to distinguish them from ions merely required to be near the polymer for the maintenance of electrical neutrality.

The chain configurations available to the polymer are assumed to be just those accessible to an otherwise similar uncharged polymer, so that the effect of the electrostatic interactions is to alter the distribution among these configurations. The chain configurations are described in the manner first proposed by Kuhn,[26] by regarding the polymer as a series of rigid links connected by universal joints. The lengths of the links are chosen so as to allow, as well as possible, for all charge-independent forces restricting the short-range flexibility of the chain. The Kuhn model makes all of the accessible configurations of equal nonelectrostatic energy, so that the charge interactions completely determine the manner in which various configurations are weighted. Thus, the charge interactions have the two following effects: they influence the equilibria among un-ionized, ionized, and bound functional groups; they affect the configurational distribution. These interactions, of course, are to be calculated, keeping in mind that the space between the charges is occupied by an ionic solution, and hence the calculation depends upon the ionic strength and upon other properties of the solution.

In accordance with the over-all approach sketched in the third preceding paragraph, one turns first to the formulation of the free energy when the mean size and degrees of ionization and binding are specified. It is convenient to begin by assigning to the polymer a specific chain configuration, and by specifying which particular functional groups are to be un-ionized, ionized, and ion paired. One may consider then the free energy of this visualizable, but unattainable, "state" of the polymer. On a completely microscopic scale, this state is really a large number of states, because the distributions of the small mobile ions of the system, the energy levels of the various species, etc., are not specified. Thus, it is relevant to consider the *free energy* of this state. Each change in the configuration or in the status of the function groups results in a new state, with its associated free energy. These states, together with their free energies, may be taken as the starting point for a statistical mechanical calculation of the over-all free energy when only the *numbers* of ionized, un-ionized, and bound groups, and the *average* configuration are specified.

The free energy of a state of the type outlined above consists only of the intrinsic free energies of the functional groups plus the electrostatic interaction free energy. The configuration of the chain does not contribute by virtue of the Kuhn chain model. The only problem is in the calculation of the interaction free energy for the charge distribution of each state. It has been shown[27] that, in spite of the presence of fixed charges, the Debye-Huckel potential can be used for this purpose. The demonstration involves approximations not much more restrictive than those of the ordinary Debye-Huckel theory.

The methods outlined above provide, at least in principle, a way to calculate the free energy of a state of specified configuration and charge distribution. At this point, mathematical complexity forces an approximation. We choose to calculate exactly for each individual state only the electrostatic interaction between nearest neighboring functional groups along the chain, and to approximate the remainder of the interaction under the assumption that the charge is distributed evenly along the chain. This approximation has the effect of preserving the strong influence that a charged group has upon the ionization and upon the binding at nearest neighboring groups, while making tractable the calculation of the less sensitive dependence on the interactions which promote configurational expansion. To be more explicit, it has just been assumed that, if the net charge is specified, the distribution of that charge along the chain is independent of the chain configuration, and that the free-energy difference among configurations is to be calculated assuming the charge to be uniformly distributed along the chain.

The remainder of the over-all free energy now separates into two parts: (a) the configurational free energy of a uniformly charged chain, and (b) the distributional free energy of a chain of charged and uncharged sites with near-neighbor interaction. The computation of (a) is essentially a biased random-walk problem, while (b) is formally equivalent to the calculation of the spin distribution in a one-dimensional Ising lattice. The details of both of these computations are described elsewhere.[25] Qualitatively, it is apparent that increasing the charge on a chain tends to restrict its configurational freedom, or to increase its configurational free energy. One would expect likewise that computation (b) might

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

result in weighting heavily only those charge distributions in which not very many more charges than necessary are nearest neighbors.

When the free energy, as computed by the methods outlined above, is minimized with respect to the number of ionized and ion-paired groups, it is found that the electrostatic interactions cause the net charge of the polymer to be much less than if no such interactions existed. This means that the free-energy increase associated with forming un-ionized groups or ion pairs is more than compensated for by the reduction in electrostatic free energy, until many more than the "normal" number of ion pairs and un-ionized groups have been created. The effect is much larger than one might suppose at first, inasmuch as the charges are frequently close enough together such that the effective dielectric constant of the region between them is far lower than that of normal water. Moreover, as the number of charged groups increases, the vast majority of the possible distributions of the charges along the chain becomes extremely unfavorable energetically, so that the free energy contains a large negative entropy term. The interaction free energy thus makes it progressively more difficult to remove successive protons from a weak polyacid, spreading the titration curve in the manner experimentally observed. When the protons finally are removed, many of them are replaced by counterions having little intrinsic affinity for ion pairing.

The configurational expansion accompanying neutralization of a polyacid is calculable by minimizing the free energy with respect to its average end-to-end distance. It is found that the binding and the reduced ionization cause the expansion as predicted for the polymer to be in much better agreement with experiment than if the binding or the reduced ionization were ignored. In addition, the amount of binding predicted to satisfy the titration and configuration properties is consistent with the amount of site binding observed in electrophoresis.[21]

The model as set forth in the preceding section may be applied to polyelectrolyte gels[28] which may be analogues of some of the functions of cell membranes. The basic concepts are the same as those already described, but the details are significantly affected by the crosslinking of the gels. The specific systems for which the theory was developed are strong electrolyte ion exchangers in equilibrium with various kinds of monatomic counter ions. Examples of such systems are the sulfonated crosslinked polystyrene exchangers in equilibrium with mixtures of alkali metal ions. There is nothing, however, in the theory restricting the analysis to this case, and the general concepts are applicable to any charged gels or membranes.

Consider, therefore, a crosslinked gel containing a number of functional groups which can be ionized (-1 charge) or which can be paired with either one of two types of counter ion to give an electrically neutral ion pair. As before, it is assumed that the intrinsic free energy associated with the formation of each type of ion pair has a value dependent upon the kind of counter ion and upon its concentration in the external solution. The ion-exchange selectivity is introduced by assuming different intrinsic free energies for the formation of different kinds of ion pairs. Exchange selectivity results only, however, when the conditions are such that significant numbers of ion pairs can be formed.

Just as for linear polyelectrolytes, it is desired to compute the free energy for a given configuration and charge distribution of the gel. As pointed out, the Debye-Hückel screened Coulomb potential may be used to calculate the electrostatic interactions. The crosslinked systems, however, present problems not encountered when dealing with linear polyelectrolytes. Many exchange resins are crosslinked sufficiently such that there are reasonably well-defined regions inside and outside the exchanger, respectively. The interior region is characterized by different ion concentrations and by the presence of enough organic matter to cause the solution to be quite different from that outside. The analysis confirming the use of the screened Coulomb potential[27] shows that the screening constant to be used must involve the proper dielectric constant for the interior region, and must recognize that the space occupied by organic matter is not accessible to the mobile ions. The bulk concentrations should be used, as the different ion concentrations within the resin are allowed for in the derivation of the screened Coulomb potential.

The large concentration differences between regions internal and external to an ion-exchange resin have led many investigators to include an "osmotic" contribution to the free energy when describing an ion-exchange resin on a phenomenological basis. The present model, however, is entirely molecular in character, and it would be incorrect to include such a contribution in addition to the direct calculation of the detailed interactions. The concentration differences are merely a consequence of the forces already considered. The point of view presented here removes a conceptual difficulty in treating resins of low crosslinking. For, if an osmotic description is used in preference to a detailed model, ambiguity arises when the degree of crosslinking is diminished to the point where the regions inside and outside the resin cease to be well defined.

Although there is in principle a means of determining the free energy at every specified configuration and charge distribution, to simplify the calculation approximations again must be made. The difference in configurational properties of linear systems and gels suggests that the approximations used here be somewhat different from those of the preceding section. In a gel, many of the spatially near-neighboring functional groups are not near neighbors along the polymer chains, so that it is definitely necessary to consider interactions

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

between groups which are not closely connected. In fairly concentrated gels, near neighbors to a given group in general are distributed so that most motions at constant gel volume do not affect the electrostatic energy seriously, with some pairs of charges approaching more closely as others are carried further apart. This is a radically different state of affairs from that of a linear polymer, where at a constant end-to-end distance the electrostatic energy depends critically upon the local coiling of the chain. It is assumed, therefore, without gross inconsistency, that all gel configurations of the same volume are electrostatically equivalent, and that the interaction between the charge state of the gel and its configuration depends only upon the gel volume. For the purpose of calculating the electrostatic interactions, the functional groups of the gel are assumed to be on a lattice whose scale is a function of the gel volume. The distributional free energy of the gel then can be calculated subject to the same sort of approximations as enters into the simple cell theories of liquids.

The calculation of the gel free energy is complete when the configurational free energy and the electrostatic free energy are each characterized as a function of gel volume. As pointed out above, the electrostatic interactions may be handled by the methods of the cell liquid theory. The analogy is exact if ion-paired groups are identified as the holes of the liquid theory, and the charges as occupied cells. The details have been given elsewhere.[28] The configurational free energy is simply that of an uncharged gel, so that the theory of rubber elasticity may be applied. We have found it convenient to use the theory in the form given by Flory.[26]

From the free-energy function determined in the manner just described, one can relate the external solution composition to the volume and to the ion pairing in a resin. The model differs from the traditional approaches in that the only source of selectivity enters through the intrinsic ion-pairing constants. It has been assumed that the size of the ions enters in no other way. Another difference is the assumption that there is an equilibrium between paired and free charged groups, with only the paired groups contributing to a selective effect. This concept leads to the conclusion that exchange selectivity is increased when conditions are such that ion pairing increases. Since the ion pairing occurs only when necessary to reduce strong electrostatic fields, one may understand why resins of high exchange capacity should be more selective than resins of lower capacity. Likewise, resins unable to reduce their electrostatic interactions through expansion because of high crosslinking are more selective than loosely crosslinked resins which easily can swell.

Flexible polyelectrolytes have been analyzed here in terms of ion-pair formation. There is some controversy on this point though the author believes the available evidence supports the concept of ion pairs. Further discussion of this topic may be found elsewhere.[29]

The preceding description of flexible polyelectrolytes may be extended easily to polyampholytes.[30] No new considerations are necessary, excepting those already discussed for rigid polyampholytes—i.e., fluctuation attraction, local charge structure and its effect on titration behavior, and possible ion binding. However, for flexible polymers, it must be remembered that the internal molecular configurations are subject to the forces arising from these phenomena, a complication absent in the case of rigid ions. A quantitative formulation of the models as discussed in this section is considered next.

III. SUMMARY OF THE MATHEMATICAL THEORY OF RIGID POLYIONS

Consider as the first and simplest case a rigid polyion of arbitrary shape in a volume v of electrolyte.[9–12] For simplicity, attention is restricted to the case where the charge on the polyion is due to the adsorption or dissociation of groups of one species of charge. The existence of adsorbed ions also implies the existence of a dissociation equilibrium which may be characterized, of course, by the change in chemical potential per ion adsorbed. If all of the ions are discharged, the Helmholtz free energy may be written

$$A_0(n)=\sum_{i=0}^{r} N_i\mu_i(0)+a_0(n)$$
$$a_0(n)=\int_s a_0^*(\nu)dS, \qquad (1)$$

where $a_0(n)$ is the change in free energy when n discharged ions are adsorbed on the polyion, $\mu_i(0)$ is the chemical potential of an ion of species i when there are no adsorbed ions, ν is the surface density of adsorbed ions, and the integration is over the surface of a macromolecule. Each element of volume-containing ions interacts not only with the charges on the polyion surface, but also with other parts of the medium external to the macromolecule. If τ_i is the chemical potential of ionic species i,

$$\tau_i=\mu_i(0)+\chi_i(1)+q_i\int_0^1 \phi_i(\lambda,n)d\lambda, \qquad (2)$$

where λ is the fractional charge on an ion (a hypothetical charging process is used to calculate the change in free energy due to electrostatic forces), $\chi_i(1)$ is the free energy arising from the interaction of the fully charged ($\lambda=1$) central ion and its ionic atmosphere, and $\phi_i(n)$ is the average electrostatic potential in the bulk of the solution at very large distances from the polyion. If the average electrostatic potential at the point r is $\Psi_i(r)$, one has

$$\Psi_i(r)=\psi_i(r)+\phi_i(n), \qquad (3)$$

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

with $\psi_i(r)$ the change in average potential experienced by ion i in being brought from the bulk of the medium to the point r. The total change in Helmholtz free energy due to the excess charge and hence excess electrostatic interaction of the double layer over that in the uniform bulk medium easily is seen to be

$$A_e=\int_0^1 d\lambda\left[\sum_0^r q_i\int_v n_i\psi_i dv+q_i\int_s \nu(s)\psi_1 dS\right], \quad (4)$$

where it is assumed that the surface charge is due to species 1. Note that Eq. (4) is of the simple form, charge multiplied by potential, with $n_i(r)$ the volume density of ions of species i at r and $\nu(s)$ the corresponding surface-charge density. The total free energy is then (solvent is represented as species zero)

$$A(n)=N_0\mu_0(0)+\sum_1^r N_i\tau_i$$
$$+a_0(n)-n\chi_1{}^{\text{bulk}}(1)+n\chi_1{}^{\text{surface}}(1)+A_e(n). \quad (5)$$

To determine the equilibrium number of adsorbed ions, the free energy represented in Eq. (5) must be minimized with respect to n, giving the result

$$\chi_1{}^{\text{bulk}}(1)-\chi_1{}^{\text{surface}}(1)=\frac{\partial a_0}{\partial n}+\frac{\partial A_e}{\partial n}$$
$$+\frac{\partial}{\partial n}\sum_1^r q_i\int_0^1[\phi(\lambda,n)-\phi(\lambda,0)]d\lambda, \quad (6)$$

with the last term accounting for the change in composition and hence mean potential of the bulk medium when n ions are adsorbed.

Consider now the interaction of two polyions, fixed in position a distance R apart. The free energy of this system differs from the preceding only in that the relevant mean potentials must be obtained as a function of R. One may write

$$A(n,R)=N_0\mu_0(0)+\sum_1^r N_i\tau_i$$
$$+2a_0(n)-2n(\chi_1{}^{\text{bulk}}(1)-\chi_1{}^{\text{surface}}(1))+A_e(n,R), \quad (7)$$

where $A_e(n,R)$ is to be computed now for two double layers, and the force between the macroions is

$$\left(\frac{\partial A(n,R)}{\partial R}\right)_{n=n_{\text{eq}}}=\frac{\partial}{\partial R}\Big[A_e(n,R)$$
$$+\sum_1^r q_i\int_0^1[\phi(\lambda,n)-\phi(\lambda,0)]d\lambda\Big]_{n=n_{\text{eq}}} \quad (8)$$

The derivatives are to be evaluated along the equilibrium-adsorption isotherm which may be a function of the interparticle separation. As before, the second term on the right-hand side of Eq. (8) is the correction arising from the change in mutual energy of the small ions of the solution. To implement this general result, the right-hand side of Eq. (8) must be evaluated. Solution of the linearized Poisson-Boltzmann equation gives the following complicated results[10–12]: The average potential of each particle may be shown to be

$$\langle\Psi_1\rangle=\frac{nq_1}{Da(1+\kappa a)}\left\{1+\frac{\exp\left[-\kappa a\left(\frac{R}{a}-2\right)\right]}{(R/a)(1+\kappa a)}\right\}, \quad (9)$$

where a is the radius of a spherical macroion and κ the usual screening parameter. The interaction energy has two terms,

$$\delta W_1(\lambda,R)=\frac{\lambda q_1}{Da(1+\lambda\kappa a)^2}\frac{\exp\left[-\lambda\kappa a\left(\frac{R}{a}-2\right)\right]}{(R/a)} \quad (10)$$

and

$$\delta W_2(R)=-\frac{5\kappa_0 q_1{}^2 q_2}{4D^2 akT}\cdot\frac{a}{R(\kappa a)^5}\left[\frac{1}{\left(\frac{R}{a}-2\right)}\left\{\frac{R}{a}+1-\frac{2}{\left(\frac{R}{a}-2\right)}+\frac{2}{\left(\frac{R}{a}-2\right)^2}\right\}\right]$$
$$-\left(\frac{R}{a}+2\right)\exp\left(\frac{R}{a}-2\right)\left\{Ei\left(\frac{R}{a}-2\right)-Ei\left[\left(\frac{R}{a}-2\right)(1+\kappa a)\right]\right\}$$
$$-\frac{\exp\left[-\kappa a\left(\frac{R}{a}-2\right)\right]}{(R/a)-2}\left\{\frac{2}{[(R/a)-2]^2}+\frac{2(\kappa a+1)}{[(R/a)-2]}+\frac{[(R/a)-2]}{1+\kappa a}+(\kappa a-1)^2+2+\frac{(\kappa a)^5\left(\frac{R}{a}-2\right)}{5(1+\kappa a)^5}\right\}, \quad (11)$$

where

$$Ei(x)=\int_x^\infty\frac{e^{-u}}{u}du. \quad (12)$$

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

For very large R,

$$\delta W_2(R) \sim -\frac{30\kappa_0 q_1^2 q_2}{D^2 akT(\kappa a)^5}\left(\frac{a}{R}\right)^6 \qquad (13)$$

and both forces are repulsive.

Direct numerical integration of the complete Poisson-Boltzmann equation leads to potential distributions differing quantitatively but not qualitatively from that in Eqs. (10) and (11). Graphical and numerical results can be found in the paper by Hoskin and Levine.[10–12]

At most distances, $\delta W_1 \gg \delta W_2$, so that the dominant repulsive potential is given by Eq. (10), but at very large distances the term of Eq. (12) is dominant. The potential, Eq. (10), was considered by Verwey and Overbeek[9] and shown to lead to quantitative description of the stability of lyophobic colloids with respect to changes in ionic strength, coagulation, etc. For further details, the reader is referred to their book.

A rather different study of the nature of the repulsive interactions was undertaken by Steiner.[31] Because of the strength of the repulsive forces, there is considerable ordering of the polyions in solution. Consequently, by inversion of the angular distribution of scattered light from suitably chosen colloidal solutions (i.e., those that exhibit diffraction due to interparticle interference), the radial-distribution function may be determined experimentally. Steiner has carried out such a program for AgI sols in which the charge is due to the adsorption of I^-. Figure 1 shows the radial-distribution function deduced by Steiner (reproduced from his Ph.D. dissertation). The structure of the solution alters with dilution and ionic strength in the expected manner. It is found that the strong repulsive forces prevent the particles from approaching closer than 1500 A at low ionic strengths. The maximum in the distribution function occurs at about 3000 A, so the long range of the repulsive force is not unreasonable. Further, the shift in position of the maximum in the distribution function is approximately of the magnitude expected from the concentration change, and the addition of a quantity of bivalent cation produces an effect roughly equivalent to that of one hundred times its concentration of monovalent cation, in accordance with the Schulze-Hardy rule.

Quantitative examination of the data reveals, however, deviations from the theoretical predictions. From the shape of the radial-distribution function, it is known that the forces are repulsive. To determine the nature of these forces, one may plot $\log(V/kT)$ against $\log R$ (appropriate to a repulsion proportional to R^{-n}) and $\log(RV)$ against R (appropriate to a repulsion of the form e^{-kr}/R). Neither one of these plots is at all linear, nor does either one appear to be rectifiable by the addition or subtraction of linear parts. It is possible, even probable, that this deviation is due to the known polydispersity of size among the colloidal particles. Certainly, though the results are fragmentary, the repulsive forces calculated by Levine and by Verwey and Overbeek are not of the form found in experiments of Steiner. A discrepancy other than that due to polydispersity would be expected to show up in the radial-distribution function owing in part to the linearization of the Poisson-Boltzmann equation. Still another possible source of error is the neglect of third-body polyions in the calculation of the potential of mean force, since, in the solutions investigated, the range of the force is great enough to involve simultaneous interaction of more than two particles. Despite the apparent sophistication of the mathematical treatment, it is evident that much work remains to be done before quantitative agreement with experiment is attained.

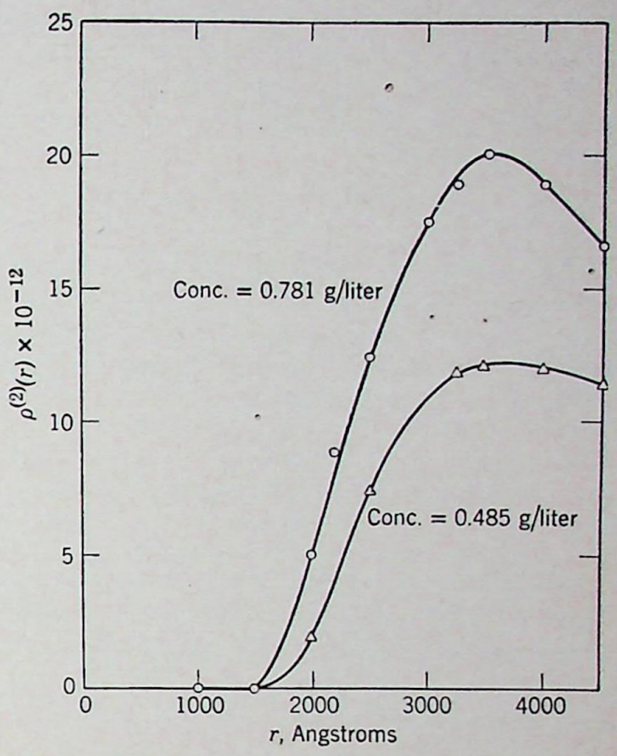


FIG. 1. A typical radial-distribution function for AgI sol.

IV. RIGID POLYAMPHOLYTES[14–16]

Polyampholytes contain both positive and negative charges on the surface of the macromolecule. It is fairly obvious that, far from the isoelectric point when charge of only one sign is present, the considerations presented in the previous section suffice to describe the interactions. At the isoelectric point where the average net charge is zero, conditions are rather different.

Consider two polyampholytes with their centers of mass separated by the distance R. The interaction between the two molecules may be written

$$V = \sum_{i=1}^{z_1}\sum_{j=1}^{z_2} \frac{\eta_i^{(1)}\eta_j^{(2)} q^2}{DR_{ij}^{(12)}} \frac{\exp[-\kappa(R_{ij}^{(12)} - a_{12})]}{1+\kappa a_{12}}, \qquad (14)$$

where there are Z_1 and Z_2 sites on polyions one and two, respectively, q is the magnitude of the charge, $R_{ij}^{(12)}$ is the separation of site i on molecule one from site j on

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

molecule 2, and $\eta_i^{(1)}$ and $\eta_j^{(2)}$ are occupation variables which can have the values zero for an uncharged site and plus or minus unity for positive and negative charges on the specified site. Now the potential of mean force is defined by

$$\exp(-W/kT)=\langle\exp(-V/kT)\rangle, \tag{15}$$

where the average is taken as unweighted and with molecules one and two fixed. It may be shown readily that the potential of mean force may be expressed in terms of unweighted averages of the potential energy V.[32] We use the truncated form,

$$W(R)=\langle V\rangle-\frac{1}{2kT}(\langle V^2\rangle-\langle V\rangle^2), \tag{16}$$

correct to terms of order $(kT)^{-1}$. At the isoelectric point, the average net charge is zero and the first term vanishes, as do all powers of $\langle V\rangle$. Now define

$$\begin{aligned} q_i^{(1)}&=q\eta_i^{(1)}=\langle q_i^{(1)}\rangle+\Delta q_i^{(1)}\\ \Delta q_i^{(1)}&=q[\eta_i^{(1)}-\langle\eta_i^{(1)}\rangle]=q\Delta\eta_i^{(1)}. \end{aligned} \tag{17}$$

The use of Eq. (14) in Eq. (16) gives

$$W(R)=-\frac{q^4}{2kTD^2R^2}\sum_{i=1}^{z_1}\sum_{l=1}^{z_1}\sum_{k=1}^{z_2}\sum_{s=1}^{z_2}\langle\eta_i^{(1)}\eta_l^{(1)}\rangle\langle\eta_k^{(2)}\eta_s^{(2)}\rangle \times\frac{e^{-2\kappa(R-a)}}{(1+\kappa a)^2}\left\langle\frac{R^2}{R_{ik}^{(12)}R_{ls}^{(12)}}\right\rangle, \tag{18}$$

where the last factor accounts for differences in charge-charge separation due to distribution of charges on the surface of the molecule. With the substitution of Eq. (17) into Eq. (18), the potential of mean force becomes (setting $\langle R^2/R_{ik}^{(12)}R_{ls}^{(12)}\rangle=1$, valid at large separations)

$$W(R)=-\frac{q^4}{2kTD^2R^2}\sum_{i,l=1}^{z_1}\sum_{k,s=1}^{z_2}[\langle\eta_i^{(1)}\eta_l^{(1)}\rangle-\langle\eta_i^{(1)}\rangle\langle\eta_l^{(1)}\rangle] \times[\langle\eta_k^{(2)}\eta_s^{(2)}\rangle-\langle\eta_k^{(2)}\rangle\langle\eta_s^{(2)}\rangle]\frac{e^{-2\kappa(R-a_{12})}}{(1+\kappa a_{12})^2}. \tag{19}$$

This term represents the interactions between the fluctuating charges and fluctuating multipoles of both molecules. The charge fluctuations may be obtained easily by differentiation of the Grand Partition Function, as is seen later. To calculate the excess chemical potential, one uses the relation

$$\frac{100\bar{M}_1}{NkT}\left(\frac{\partial\mu_1^e}{\partial c_1}\right)_{T,p,\kappa_0}=4\pi\left[\int_0^\infty R^2[g_{10}(R)-1]dR-\int^\infty R^2[g_{11}(R)-1]dR\right], \tag{20}$$

where, in the usual notation,

$$\begin{aligned} \mu_1&=\mu_1^0(T,p)+kT\ln c_1+\mu_1^e\\ \mu_1^0(T,p)&=\lim_{c\to0}[\mu_1-kT\ln c_1]\\ g_{10}(R)&=\exp\left[-\frac{W^{(10)}}{kT}\right]\\ g_{11}(R)&=\exp\left[-\frac{W^{(11)}}{kT}\right] \end{aligned} \tag{21}$$

with $W^{(10)}$ and $W^{(11)}$ the potentials of mean force between a macromolecule and a solvent molecule and between two macromolecules, respectively. The dimensions of a solvent molecule are small relative to those of the macromolecule in almost all cases, and the solvent plays the role of a dielectric continuum. One finds, therefore, after substitution and integration, that

$$\frac{\mu_1^e}{kT}=\frac{c_1}{100M_1}\left[\frac{7\pi Na_{12}^3}{6}-\frac{\pi Nq^4\langle\Delta\eta_i^{(1)2}\rangle}{(DkT)^2\kappa_0(1+\kappa_0a_{12})^2}\right], \tag{22}$$

where the macroion contribution to the ionic strength is neglected.

At high salt concentrations, κ_0 is large and the second term is negligible relative to the first term. Under these conditions, the excess chemical potential is positive, representing the net repulsive forces which result in an excluded volume. When the salt concentration is very low, the second term dominates and the excess chemical potential is negative. Careful light-scattering experiments have verified Eq. (22) quantitatively.[33] The magnitude of the fluctuating charge may be obtained independently from the titration curve of the polyampholyte and is in complete agreement with the values determined from the independent light-scattering experiments.[34]

A very interesting application of this theory has been presented by Kirkwood[17] in a discussion of enzyme kinetics. Suppose the enzymatic reaction can be described by the Michaelis-Menten relation

$$-\frac{d[S]}{dt}=\frac{k_3[E][S]}{K_m+[S]} \tag{23}$$

in which $[S]$ and $[E]$ are the substrate and enzyme concentrations, K_m the Michaelis-Menten constant, and k_3 the intrinsic rate constant for the enzymatic reaction. Now the potential of mean force is the work required to bring a pair of molecules from infinite separation to the distance R. If the standard free energy of formation of a complex is ΔG^0, then this is equal to the potential of mean force at contact; i.e.,

$$\Delta G^0=W(a), \tag{24}$$

and, therefore,

$$kT\ln K=-W(a), \tag{25}$$

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

where K is the equilibrium constant for the formation of the complex. Let the complex now undergo reaction through the intermediary of a transition state with free energy of activation $\Delta G^{\pm}$. Thus,

$$\Delta G^{\pm} = W^{\pm}(a^{\pm}) - W(a). \tag{26}$$

Suppose the substrate molecule is specifically bound to the active site of the enzyme by local forces with potential W_0, in which the protein does *not* participate. In the activated state, the potential is denoted by $W_0^{\pm}$, the subscript zero again indicating that protein forces are omitted. Under the influence of the protein, the local forces are different and the corresponding potentials are denoted W' and $W^{\pm}$. If K_m and K_m^0 and k_3 and k_3^0 are the Michaelis-Menten constants and intrinsic rates of reaction with and without protein participation, then

$$kT \ln \frac{K_m}{K_m^0} = W' - W_0,$$
$$kT \ln \frac{k_3}{k_3^0} = \Delta W_0^{\pm} - \Delta W^{\pm}, \tag{27}$$

where the intrinsic rate of decomposition of ES into reactants is assumed to be small relative to k_3.

If the protein contains Z groups with acid dissociation constants K_i, and if the substrate molecule has dipole moments μ' and $\mu^{\pm}$ in the normal and activated states, then, neglecting all forces except those arising from charge fluctuations leads to

$$W - W_0 = -\frac{1}{2kT} \sum_1^z \frac{K_i[H^+]}{([H^+]+K_i)^2} \frac{q^2\mu^2 \cos^2\gamma_i}{D_e^2 R_i^4}, \tag{28}$$

where γ_i is the angle between the dipole moment of the substrate molecule and the radius vector $\mathbf{R}_i$ from site i, D_e is the effective dielectric constant, and the first factor arises from the evaluation of $\langle \eta_i^2 \rangle$ as is seen in a subsequent section.[30] Note that electrostatic interactions between charges on the protein again are neglected.

Since the fluctuation potentials fall off as R^{-4}, one may simplify Eq. (28) by assuming that only nearest neighboring groups are effective. If there are Z_α nearest neighboring groups with dissociation constants K_α,

$$W - W_0 = -\frac{Z_\alpha q^2 \mu^2}{4D_e^2 r_\alpha^4 kT} \frac{K_\alpha[H^+]}{([H^+]+K_\alpha)^2}$$
$$r_\alpha^{-4} = \frac{2}{\nu_\alpha} \sum \cos^2\gamma_i / R_i^4. \tag{29}$$

From Eqs. (29) and (27),

$$\log \frac{k_3}{k_3^0} = \frac{Z_\alpha q^2 (\mu^{\pm 2} - \mu'^2)}{4D_e^2 k^2 T^2 r_\alpha^4} \frac{K_\alpha[H^+]}{([H^+]+K_\alpha)^2}. \tag{30}$$

Note that, for this simplified model, k_3 has a maximum when the pH is equal to pK_α. Further, the pH dependence predicted by the model is a symmetric bell-shaped curve. If the electrostatic interactions between charges on the surface of the protein are accounted for, then the curve may be skewed since the proton population on the molecular surface, and hence the free energy of activation, are not symmetric functions of the pH. Kirkwood has made a crude comparison of the theory with some experimental data of Bergmann and Fruton on the effect of pH on the pepsin hydrolysis of carbobenzoxy-1-glutamyl-1-tyrosine. The maximum rate occurs at pH 4, thereby identifying the participating groups as carboxylates. With an estimated dielectric constant and an estimated dipole moment difference, the fall in rate to one-half of its value in an interval of one pH unit is accounted for if there are about ten carboxyl groups at an average distance of 5 A from the adsorption site. This picture is qualitatively reasonable. It should be stressed that we have considered only one of many possible contributory forces operative in enzyme reactions and it is not to be implied that the charge-fluctuation mechanism is the only or even the most important of the possibilities. The charge-fluctuation mechanism, however, does provide a physical picture in qualitative accord with experiment.

It should be apparent by this time that the calculation of the properties of an isolated polyion eventually requires the computation of the mutual electrostatic energy of the charges on the molecule. Until very recently, it has been customary to calculate this electrostatic free energy for spherical macroions by assuming the charge to be uniformly distributed over the surface, and the macromolecule to be impenetrable by the small ions of the bulk medium. The result obtained is,

$$A_{\text{unif}}^{\text{sphere}} = \frac{q^2 Z^2}{2Da}\left[1 - \frac{\kappa a}{1+\kappa a}\right], \tag{31}$$

where Zq is the net charge on the ion. Aside from the questions of penetrability which introduce no basically new features,[18] the assumption of a uniform charge distribution is grossly inaccurate. The calculation of the electrostatic interaction energy for an arbitrary distribution of positive and negative charges is extremely difficult. Basically, the method consists in the direct counting of all configurations with given total charge and weighting each configuration with the appropriate Boltzmann factor. Fortunately, for small numbers of charges, there are many configurations of equal energy, and this degeneracy simplifies the calculation. With the energy zero taken as the completely discharged protein, the work required to charge a polyampholyte sphere

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

may be shown to be[14,15]

$$A^{\text{sphere}}_{\text{discrete}}=\frac{q^2}{2a}\sum_{k=1}^{z}\sum_{l=1}^{z}\eta_k\eta_l(\mathfrak{a}_{kl}-B_{kl}) -\frac{q^2}{2a}\sum_{k=1}^{z}\sum_{l=1}^{z}\eta_k\eta_l C_{kl}, \quad (32)$$

where

$$\mathfrak{a}_{kl}=\frac{a}{D_i r_{kl}},$$

$$B_{kl}=\frac{1}{D_i}\sum_{n=0}^{\infty}\frac{(n+1)(D-D_i)}{(n+1)D+nD_i}\rho_{kl}{}^n P_n(\cos\vartheta_{kl}),$$

$$C_{kl}=\frac{1}{D}\left[\frac{\kappa a}{1+\kappa a}+\sum_{n=1}^{\infty}\frac{2n+1}{2n-1}\left\{\frac{D}{(n+1)D+nD_i}\right\}^2 \times\frac{(\kappa a)^2\rho_{kl}{}^n P_n(\cos\vartheta_{kl})}{\frac{K_{n+1}}{K_{n-1}}+\frac{n(D-D_i)}{(n+1)D+nD_i}\cdot\frac{(\kappa a)^2}{4n^2-1}}\right], \quad (33)$$

$$\rho_{kl}=r_k r_l/a^2,$$

$$K_n(\kappa a)=\sum_{s=0}^{n}\frac{2^s n!(2n-s)!}{s!(2n)!(n-s)!}(\kappa a)^s,$$

with $P_n(\cos\vartheta_{kl})$ as the Legendre polynomial of order n, r_k, r_l, r_{kl}, and ϑ_{kl} are, respectively, the distances of charges k and l from the center of the polyion, the distance between charges k and l, and the angle between the vectors $\mathbf{r}_k$ and $\mathbf{r}_l$. To this electrostatic energy must be added the chemical-potential change due to the intrinsic change in free energy on ionization. This total free energy then represents a protein with specified charge distribution, and to obtain the relevant macroscopic free energy, an average over the charge distribution must be performed. The theory has been applied to models of proteins in which the charges were variously located at the vertices of a cube and dodecahedron.[14,15] The results of these laborious calculations may be summarized as follows. The titration curve depends markedly upon the depth of the charges below the surface and upon the distribution of charges. As expected, the interaction energy is smallest for a regular, uniform distribution, larger for a random distribution, and very large if charges are crowded at one end of the molecule and the charge distribution made nonuniform. Arguments can be offered which suggest that the charge groups in a protein are uniformly about one angstrom below the surface.[14,15] All deviations from the electrostatic energy given in Eq. (31) can be accounted for in terms of the charge distribution on the molecule. Great care must be used in interpreting variations in titration curve with configurational changes in the macromolecule.

V. FLEXIBLE POLYELECTROLYTES

In turning to a consideration of the properties of flexible polyelectrolytes, a complicating feature appears. This is the change in mean dimensions of the molecule when the ionic strength or charge density is altered. As stated earlier, the change in configuration is caused by a readjustment between the contractile forces due to the Brownian motion of the stretched chain, a force of entropic origin, and the repulsive forces due to the electric charges. Using the procedures discussed in the preceding sections, it may be shown that the Helmholtz free energy of an independent polyion immersed in a medium of small ions is of the following form[25]:

$$A=A_1+A_2+A_3+A_4 \quad (34)$$

with

$$A_1=Z\alpha\mu_{H^+}+Z\alpha(1-f)\mu_{C^+}$$

$$A_2=-ZkT\ln\tfrac{1}{2}\Big(\lambda e^{-\chi/kT}+1 +\Big[(\lambda e^{-\chi/kT}+1)^2-4\lambda(e^{-\chi/kT}-1) +Z\alpha' kT\ln\lambda-\frac{q^2\kappa Z\alpha'}{2D}\Big]^{\frac{1}{2}}\Big) \quad (35)$$

$$A_3=-Z\alpha(kT\ln K_a+\mu_{H^+}{}^0) +Z\alpha f(kT\ln K_s{}^0+\mu_{C^+}{}^0)-TS_m$$

$$A_4=-kT\ln Q^0-Z_k kT\ln\frac{\int e^{-u(\gamma)/kT}d\Omega}{4\pi}.$$

In Eqs. (35), α' is the degree of ionization of the macromolecule and f the fraction of dissociated sites occupied by bound-ion pairs. The degree of neutralization then is related to the degree of ionization by the conservation condition $\alpha'=\alpha(1-f)$, and the fraction of bound sites referred to the total number of sites of the polymer will be αf. A_1 describes the free energy of the free hydrogen ions and counterions eligible to take part in binding and neutralization phenomena. A_2 represents the electrical free energy of the net charge of the polyion, regarding the ion pairs as uncharged sites and thus includes the interaction energy between portions of the net charge of the polymer and the entropy associated with the mixture of charged and uncharged sites along the polymer chain. A_3 includes the chemical free energy associated with the state of ionization and binding of the polymer, computed from a reference state in which the polyion is in its completely undissociated form. Also included in A_3 is the free energy of mixing of the bound ion pairs and un-ionized groups among the uncharged sites of the chain. Finally, the last contribution, A_4, is the intrinsic free energy of the polymeric skeleton,

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

a function of its configuration. Of the remaining undefined symbols, S_m is the entropy of mixing uncharged ion-paired and un-ionized groups, K_a and K_s^0 are the intrinsic dissociation constants of the acid- and ion-paired groups, Z_k is the number of Kuhn statistical elements in the randomly coiled chain, and $u(\gamma)$ is the interaction potential between neighboring statistical elements making an angle γ at their junction.

From Eqs. (34) and (35), the equilibrium root-mean-square end-to-end separation of the chain and the extent of counterion binding may be evaluated by straightforward minimization. The agreement between experiment and theory is semiquantitative, as can be seen in Figs. 2 and 3 which show the amount of binding and the expansion of sodium polyacrylate and sodium carboxymethyl cellulose, respectively. Lengthy discussion of the suitability of the ion-pairing concept can be found elsewhere, and space limitations prevent any amplification of these arguments. Suffice it to say that this author believes the concept to be useful and at least partially supported by experiment, whereas other explanations (such as ion trapping in the region where $e\psi/kT>1$) of the intimate association of counterions, with the polyion cannot explain all the relevant data.[29]

It is apparent that the permeability of the open network representing the expanded polymer coil permits the penetration of electrolyte and the consequent reduction of the electrostatic interactions between polyions. Several approaches[27,35–37] to the solution of the Poisson-Boltzmann equation have been published. Despite widely differing approximations, the conclusion of all investigators is that the domain occupied by the polyion is essentially neutral and polyion-polyion interactions, therefore, are dominated by the excluded volume-type interactions characteristic of neutral polymers. This is in complete agreement with the data of Schneider and Doty.[38]

From the microscopic point of view, there is a very important difference between solutions of polyions and solutions of ordinary electrolytes. In an ordinary electrolyte solution, the medium is essentially homogeneous, and the work required to bring an ion from infinity to **r** does not depend upon **r** (always neglecting boundary effects). In a solution of macroions, the internal region of any polyion is, roughly speaking, a solution of much higher ionic strength than the external medium. Thus, the work required to bring a charge to the point **r** depends upon whether **r** is within or without the volume occupied by the macroion. The extra work required in this case corresponds exactly to the difference in ion-atmosphere interaction corresponding to the two different concentrations of electrolyte.[27] Because of counterion binding, the potential inside a polyion is actually much smaller than would appear to be the case. In fact, marked deviations from a linearized Poisson-Boltzmann equation are expected only in the vicinity of the unit charges of the polymer, and these deviations are no more serious than those in solutions of small electrolytes. Solution of the linearized Poisson-Boltzmann equation (with ion pairing) shows that the internal and external interactions are differently screened, corresponding to the crude picture of a drop of concentrated electrolyte suspended in dilute electrolyte. Of course, when the charge density is small, or the polymer highly extended, the different screening lengths approach equality. Nevertheless, there are systems, such as polyelectrolyte gels, for which the difference is of great importance. The conclusions sketched above also may be arrived at from the diametrically opposed assumption that, to a first approximation, the polymeric volume is electrically neutral, and from calculating the small potential difference between the interior and the bulk medium from a Donnan equilibrium condition. That the same conclusions are reached when two such widely differing approximations are used lends support to the conclusions as general results.

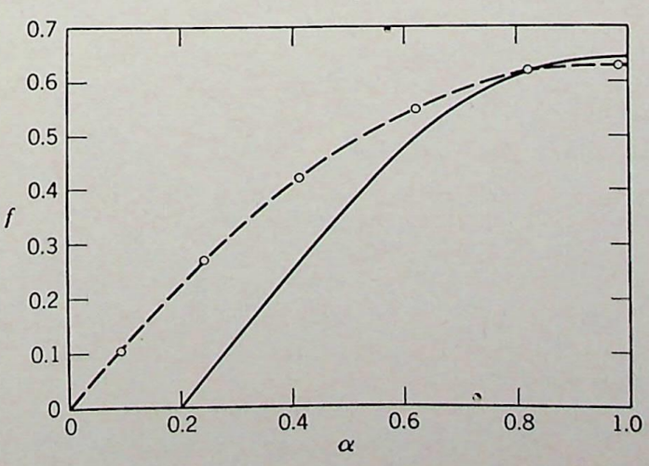


FIG. 2. Fraction of counterions bound to sodium polyacrylate as a function of the degree of neutralization. Solid line is theoretical curve. The dashed curve is from the experimental data of Huizenga, Grieger, and Wall.[21]

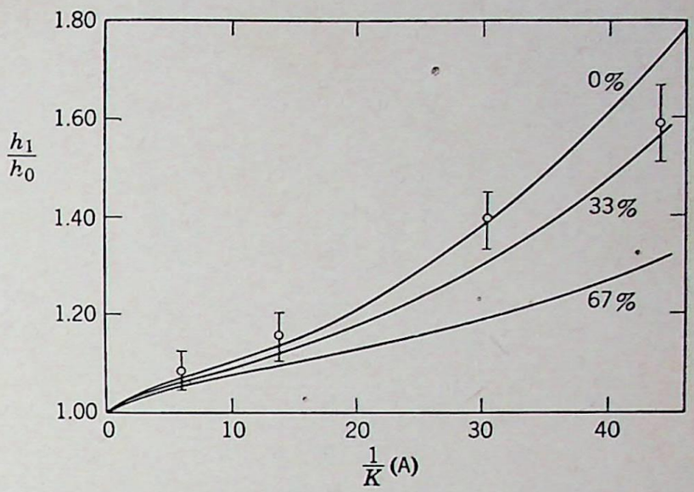


FIG. 3. Expansion of sodium carboxymethyl cellulose as a function of ionic strength. The curves are calculated assuming the amounts of counterion binding indicated. The circles are the experimental points of Schneider and Doty.[38]

While the forces which are operative in polyampho-

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

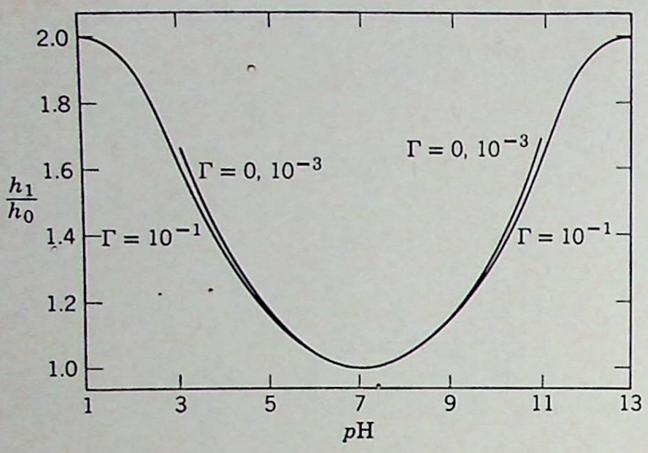


FIG. 4. Expansion as a function of pH for several ionic strengths for a regularly alternating copolymer; $K_1^0 = 10^{-5}$, $K_2^0 = 10^{-9}$, $K_3^0 = K_4^0 = 10$, charge spacings 6 and 8 A for nearest and next nearest neighboring groups. For further details, see reference 20.

lyte systems are qualitatively no different from those in purely acidic or basic polyelectrolytes, significant new problems arise because of the inhomogeneity of the polymer skeleton. Obviously, far from the isoelectric point, the polyampholyte behaves as an ordinary polyelectrolyte. On the other hand, in the nearly isoelectric region where appreciable numbers of positive and negative charges can coexist, these complications dominate the situation.

One may proceed[30] to calculate the titration curve and other thermodynamic properties by exactly the same technique as used before, taking cognizance of both the existence of two types of charge and the skeletal distribution. At the isoelectric point, it is no longer true that all sites capable of bearing a charge are the same. Even though there is no net charge on the polymer, forces acting between the elements may arise from the two following distinct sources: (a) fluctuations of the charge of each statistical element about its mean value; (b) correlations in charge distribution because of the polymerization statistics.

From the properties of the Grand Partition Function, it may be readily shown that

$$\langle (q-\bar{q})^2 \rangle = -\frac{1}{2.303 Z_\kappa} \frac{1-\alpha_1'}{1-\alpha_1} \left(\frac{\partial \bar{Q}}{\partial p H} \right)_{a_{x^+}, a_{y^-}} \tag{36}$$

for the mean-square charge on a statistical element due to fluctuations and for the hydrogen-ion activity at the isoelectric point,

$$a_{H^+} = \left[\frac{K_1^0 K_2^0}{1 + \frac{a_{x^+}}{a_{H^+}} \frac{K_1^0}{K_3^0} - \frac{a_{y^-}}{a_{H^+}} \frac{K_1^0}{K_4^0}} \right]^{\frac{1}{2}}. \tag{37}$$

The evaluation of the degree of dissociation of acid groups α_1, of binding to acid groups, $\alpha_1' = \alpha_1(1-f)$, and the total charge $\bar{Q}$, requires specific consideration of skeletal structure. The dissociation constants K_1^0, K_2^0, K_3^0, K_4^0 are for the acid, base, acid-counterion, and base-counterion pairs.

Consider the simplest possible case, when the acid and base groups alternate regularly along the polymer skeleton. It is found that:

(a) The counterion binding does not cause a large change in the titration curve near the isoelectric point. This is due to the stabilizing influence of adjacent opposite charges. When the charge is predominantly of one sign, the polyampholyte behaves essentially like a one-component polyion.

(b) The expansion of the polyion is not strongly dependent upon the ionic strength at a given pH. This is due to the fact that as the ionic strength increases, tending to weaken the force expanding the polymer, the charge which can be supported on it also increases, since the group interactions are decreased. This latter tends to counterbalance the first effect.

(c) Since the binding of counterions is small, the effective net charge is essentially identical with the number of molecules of acid or base added to the solution. The effects of added salts are to shift the pH at any given degree of neutralization towards the isoelectric point, creating a curve characteristic of weaker acidic and basic groups. The primary source of this shift in the titration curve is due to electrostatic interaction among segments of the polymer and not to binding. This effect is, of course, similar to the observed effects in proteins.

(d) Effects of correlations due to polymerization statistics vanish identically, and those due to charge fluctuations are very small. These results are depicted in Figs. 4–6.

As a more general case, consider an equimolar polyampholyte, but with statistically distributed monomeric groups. This is a much more difficult problem. To

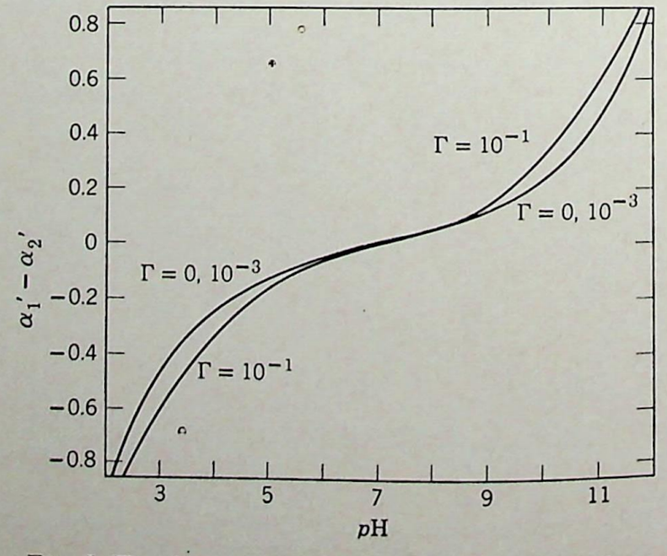


FIG. 5. Titration curve for the same regularly alternating polyampholyte as in Fig. 4.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

proceed, the polymer is characterized by the number of groups of a given type which occur in a row. It must be recognized that those charge-bearing groups adjacent to groups of the same type have a tendency to remain uncharged as compared to groups surrounded by oppositely charged sites. If the polymer is regarded as a succession of acidic and basic groups, it is clear that, near the isoelectric point, almost all of the end groups are charged. Thus, the interior groups of the sequences of monomers may be regarded as independent to a good approximation. The problem now has been reduced to a combinatorial question of how many ways one can distribute a certain number of charges over the polymer.

The results of this computation are as follows:

(a) The titration curve has a smaller variation near the isoelectric point in *p*H with changes in the net charge of the polymer than has the corresponding regularly alternating polyampholyte. This means that the slope of the titration curve at the isoelectric point is steeper, and that the effect is due to the considerable number of charges situated in the interior of sequences and adjacent to charges of the same sign. These are, therefore, relatively loosely bound and a much smaller change in the chemical potential of the hydrogen ion results from their removal than from the removal of an equal number of the charges of a regularly alternating polyampholyte. The inverse statement is perhaps clearer; i.e., it takes a smaller change in *p*H to effect a change in charge for the randomly distributed monomers.

(b) A wide range of behavior is possible due to variations in structure. One obvious extreme is the block copolymer with all acid groups at one end and base groups at the other end of the molecule.

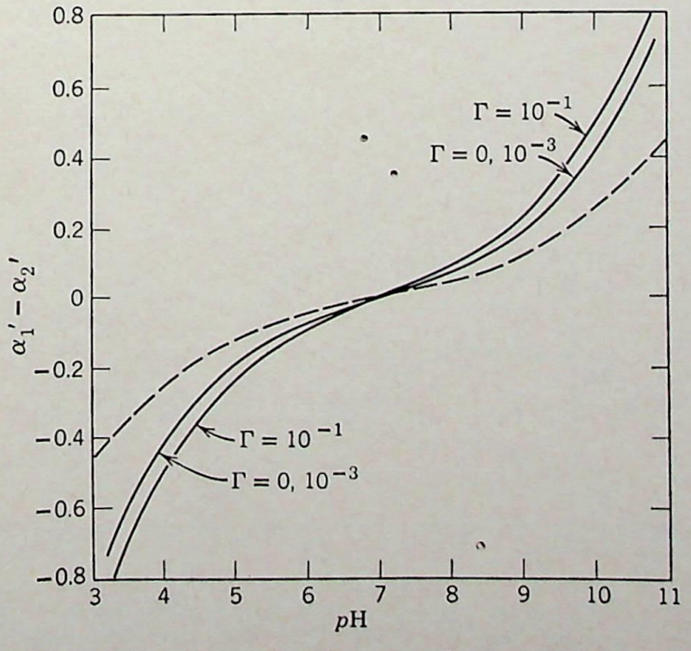


FIG. 6. Comparison of titration curves for a regularly alternating (broken line) and a randomly distributed (solid line) copolymer. The constants are the same as in Fig. 4.

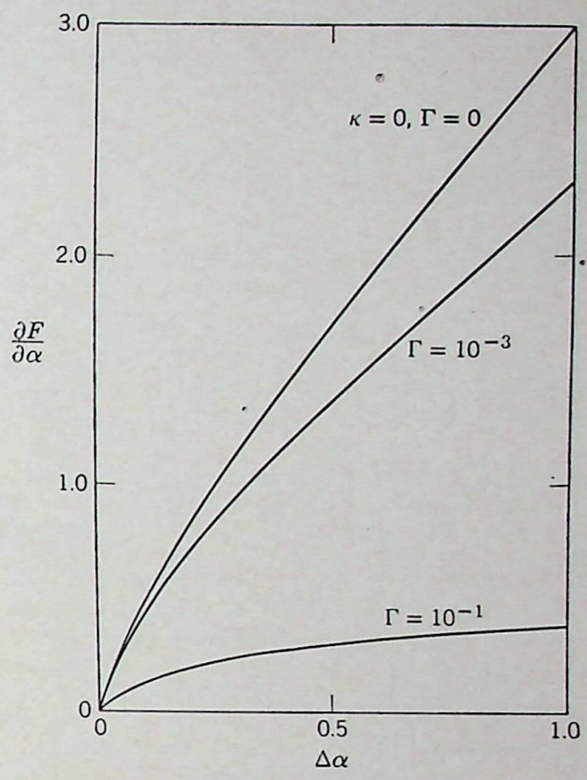


FIG. 7. Rate of change of expansion free energy with degree of neutralization as a function of net charge. The copolymer is the same as in Fig. 4.

(c) An increase in ionic strength, resulting in a decrease in interaction between charge groups, causes a shift of the titration curve in the direction of the behavior expected of independent acidic and basic groups.

(d) The configurational properties are almost identical with the corresponding regular copolymer of equivalent net segment charge.

(e) Charge fluctuation interactions are small due to the equimolarity of acidic and basic groups. This effect is large only when the numbers of one type of group exceed the other.

The properties of the regular and statistical polyampholytes are compared in Figs. 6 and 7.

VI. HELIX-COIL TRANSITIONS AS INFLUENCED BY ELECTROSTATIC FORCES[39]

The fact that the electrostatic energy of a molecule depends upon the molecular configuration suggests that electrostatic forces can be the motive source for molecular transitions. Consider a simple model wherein several isomeric states of a molecule may exist, characterized by different electrostatic energies. For simplicity, assume that the distribution of charges and uncharged sites is random in all isomeric states and that ion binding effects may be neglected. Further, for simplicity, assume the reaction to proceed in an all-or-none fashion. If the electrostatic energy is computed

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

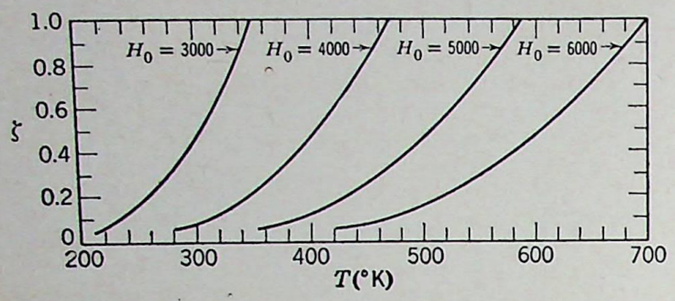


FIG. 8. Equilibrium fraction of bonds broken in DNA as a function of temperature. The broadening can be appreciably reduced by different choices of the parameters. For further details, see references 39 and 44–46.

for smeared-out charges, it is readily shown that

$$p\mathrm{H}-\log\frac{\alpha'}{1-\alpha'} = 0.868\frac{\alpha'}{Z}\frac{W_h^0+W_c^0\Theta\exp[-(\alpha')^2\Delta W^0]}{1+\exp[-(\alpha')^2\Delta W^0]} \quad (38)$$

with

$$\Delta W^0 = W_c^0 - W_h^0$$

$$\Theta = Q_c^0/Q_h^0$$

$$\frac{U_h^{\mathrm{elec}}}{ZkT} = (\alpha')^2 W_h^0 \quad (39)$$

$$\frac{U_c^{\mathrm{elec}}}{ZkT} = (\alpha')^2 W_c^0,$$

which defines the pair-interaction parameters W_h^0 and W_c^0, and Q_c^0 and Q_h^0 are the partition functions for the coil and helix without electrostatic interactions. The pair-interaction parameters are[40,41]

$$W_h^0 = \frac{Z\rho}{kT}\frac{4\pi^2 b^2 l_0}{D}\frac{K_0(\kappa b)}{\kappa b K_1(\kappa b)}$$

$$W_c^0 = \frac{Z^2q^2}{DkT}\left[\frac{6}{\kappa h_0^2} - \left(\frac{8}{3\pi}\right)^{\frac{1}{2}}\frac{h_1}{2}\left(\frac{6}{\kappa h_0^2}\right)^2\right] \quad (40)$$

with l_0 the length of a segment, b the radius of the helix, K_0 and K_1 are Bessel functions of the second kind and ρ is the charge density on the helix surface. Correspondingly, h_0 and h_1 are the root-mean-square end-to-end separations of the coil in the absence and in the presence of electrostatic interactions, respectively. Equation (38) does not apply to any real helix-coil transition because of the assumption that the reaction does not involve intermediate states, an approximation known to be poor.[42–45] For the more general case, detailed calculations for the helix-coil transition in the thermal denaturation of DNA predict equilibrium bond breakage as a function of temperature as indicated in Fig. 8.[39] These calculations are for the case when electrostatic interactions are omitted. The effect of charge-charge interaction is to lower the enthalpy required to break bonds, H_0. In Fig. 8, this corresponds to moving from right to left across the lines of constant H_0. In Fig. 9 is a titration curve for DNA at two temperatures showing that the transition occurs titrimetrically at the same place as it is observed to occur viscometrically, calorimetrically, and by several other techniques.[46] The treatment can be made much more sophisticated by the introduction of discrete charges, ion binding, differences in charge between isomeric states, and other effects previously discussed, but the qualitative results are the same. A qualitative change does occur when intermediate states between pure helix and pure random coil are considered. The reader is referred elsewhere for details.[39,44]

VII. SOME BRIEF REMARKS

Nowhere in this review are transport properties considered, a subject of great intrinsic interest. The treatment of, say, the conductance is even more complicated than of some equilibrium property due to the loss of elements of symmetry. Some progress has been made by Booth,[13] by Hermans,[47] and by Overbeek,[48] and the reader is referred to their papers.

In closing, the author wishes to emphasize that electrostatic forces only have been considered. In any

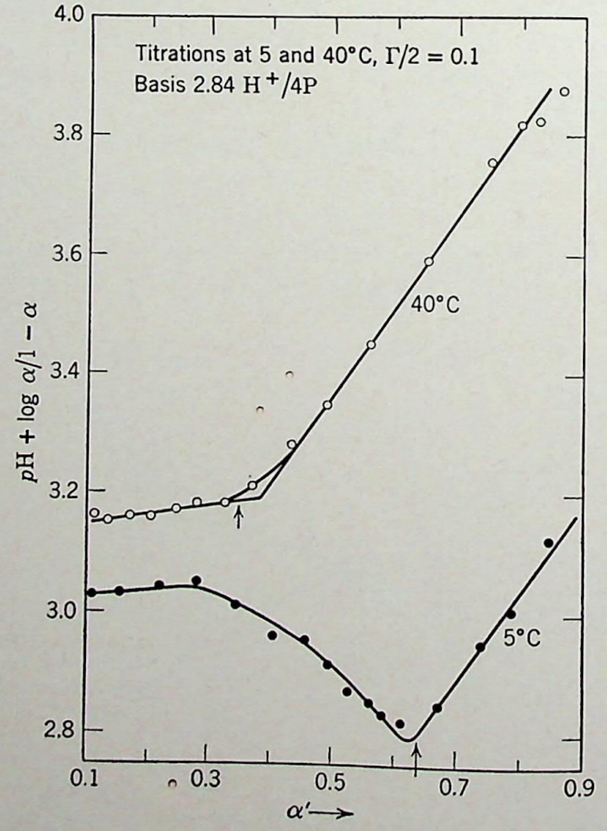


FIG. 9. Titration curve of sodium deoxypentosenucleate. The arrows indicate the points at which the helix-coil transition occurs calorimetrically and viscometrically. Note that we have plotted $p\mathrm{H}+\log\alpha/(1-\alpha)$ instead of $p\mathrm{H}-\log\alpha/(1-\alpha)$. This serves to sharpen the appearance of the transition.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

real system, many other forces are operative. To name only two, consideration must be given to hydrogen bonds and to steric hindrances. Thus, in any real system, one must account *simultaneously* for the electrostatic effects discussed and for all other relevant phenomena. Separation of these from the electrostatic part is arbitrary and not always justified. It should, therefore, be borne in mind that the applicability of the theory discussed herein to any real system must be made with forethought and with circumspection.

BIBLIOGRAPHY

[1] J. E. Mayer, J. Chem. Phys. **18**, 1426 (1950).
[2] J. G. Kirkwood and J. Poirier, J. Phys. Chem. **58**, 591 (1954).
[3] L. Onsager and S. Kim, J. Phys. Chem. **61**, 198 (1957).
[4] R. M. Fuoss and L. Onsager, J. Phys. Chem. **61**, 668 (1957).
[5] H. Falkenhagen, M. Leist, and G. Kelbg, Ann. Physik **11**, 51 (1952).
[6] E. Pitts, Proc. Roy. Soc. (London) **A217**, 43 (1953).
[7] J. G. Kirkwood, J. Chem. Phys. **14**, 51 (1946).
[8] See for example, P. Doty and G. Ehrlich, Ann. Rev. Phys. Chem. **3**, 81 (1952).
[9] E. J. Verwey and J. Th. G. Overbeek, *Theory of the Stability of Lyophobic Colloids* (Elsevier Publishing Company, New York, 1948).
[10] S. Levine, Proc. Roy. Soc. (London) **A170**, 165 (1939).
[11] S. Levine, Phil. Mag. **41**, 53 (1950).
[12] N. E. Hoskin and S. Levine, Phil. Trans. Roy. Soc. (London) **A248**, 449 (1956).
[13] F. Booth, Proc. Roy. Soc. (London) **A203**, 514, 533 (1950).
[14] J. G. Kirkwood and J. Mazur, J. Polymer Sci. **9**, 519 (1952).
[15] C. Tanford and J. G. Kirkwood, J. Am. Chem. Soc. **79**, 5333 (1957).
[16] J. G. Kirkwood and J. B. Shumaker, Proc. Nat. Acad. Sci. U. S. **38**, 863 (1952).
[17] J. G. Kirkwood, Discussions Faraday Soc. **20**, 78 (1955).
[18] C. Tanford, J. Phys. Chem. **59**, 788 (1955).
[19] W. Kuhn, O. Künzle, and A. Katchalsky, Helv. Chim. Acta **31**, 1994 (1948).
[20] J. Hermans and J. Th. G. Overbeek, Rec. trav. chim. **67**, 761 (1948).
[21] J. Huizenga, P. F. Grieger, and F. T. Wall, J. Am. Chem. Soc. **72**, 2636 (1950).
[22] P. J. Flory, J. Chem. Phys. **21**, 162 (1953).
[23] G. Kimball, M. Cutler, and H. Samelson, J. Phys. Chem. **56**, 57 (1952).
[24] F. Osawa, N. Imai, and I. Kagawa, J. Polymer Sci. **13**, 93 (1954).
[25] F. E. Harris and S. A. Rice, J. Phys. Chem. **58**, 725, 733 (1954).
[26] See, for example, P. J. Flory, *Principles of Polymer Chemistry* (Cornell University Press, Ithaca, 1953).
[27] F. E. Harris and S. A. Rice, J. Chem. Phys. **25**, 955 (1956).
[28] S. A. Rice and F. E. Harris, Z. physik. Chem. (Frankfort) **8**, 207 (1956).
[29] S. A. Rice, J. Am. Chem. Soc. **78**, 5247 (1956).
[30] S. A. Rice and F. E. Harris, J. Chem. Phys. **24**, 326, 336 (1956).
[31] R. Steiner, "Light scattering studies," Ph.D. dissertation, Department of Chemistry, Harvard University (1950).
[32] See for example, T. L. Hill, *Statistical Mechanics* (McGraw-Hill Book Company, Inc., New York, 1956).
[33] S. N. Timasheff, H. M. Dintzis, J. G. Kirkwood, and B. Coleman, J. Am. Chem. Soc. **79**, 782 (1957).
[34] S. Lowey, Ph.D. dissertation, Department of Chemistry, Yale University (1957).
[35] S. Lifson, J. Chem. Phys. **26**, 700, 727, 1356 (1957).
[36] F. T. Wall and J. Berkowitz, J. Chem. Phys. **26**, 114 (1957).
[37] M. Nagasawa and I. Kagawa, Bull. Chem. Soc. Japan. **30**, 961 (1957).
[38] N. S. Schneider and P. Doty, J. Phys. Chem. **58**, 762 (1954).
[39] S. A. Rice and A. Wada, J. Chem. Phys. **29**, 233 (1958).
[40] A. Katchalsky and S. Lifson, J. Polymer Sci. **11**, 409 (1953).
[41] T. L. Hill, Arch. Biochem. Biophys. **57**, 229 (1955).
[42] B. Zimm and J. Bragg, J. Chem. Phys. **28**, 1246 (1958).
[43] J. Gibbs and E. DiMarzio, J. Chem. Phys. **28**, 1247 (1958).
[44] B. H. Zimm, S. A. Rice, and A. Wada (to be published).
[45] S. A. Rice, A. Wada, and E. P. Geiduschek, Discussions Faraday Soc. **25**, 130 (1958).
[46] J. Sturtevant, S. A. Rice, and E. P. Geiduschek, Discussions Faraday Soc. **25**, 138 (1958).
[47] J. Hermans, J. Polymer Sci. **18**, 527 (1955).
[48] J. Th. G. Overbeek and D. Stigter, Rec. trav. chim. **75**, 543 (1956).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

8. Proteins and Their Interactions

DAVID F. WAUGH

Department of Biology, Massachusetts Institute of Technology, Cambridge 39, Massachusetts

THIS article, early in the series devoted to the soluble (globular) proteins, discusses briefly some of the properties of the native proteins, the main characteristics of protein denaturation, and a few examples illustrating specificity and complexity in the interactions between proteins. A model is first presented as a means for describing and correlating, in a general way, structure and properties.

At this point, one accepts at once the facts that globular (soluble) proteins are condensed structures, that they have many of the characteristics of molecules and, from their diffraction patterns, that portions of the main or backbone chain of atoms are arranged with considerable regularity. Of the possible arrangements for the latter, that of a helix has been most attractive. For some time, the α-helix of Pauling and Corey[1] has provided a singular foundation for discussion. Several elegant helices were discovered by Pauling and Corey by assuming (1) that the planarity of the CONH grouping around the peptide bond should be preserved, and by assuming (2) the formation of a maximal number of hydrogen bonds between CO and NH groups separated by intervening peptide groups along the main chain. In the α-helix, a compact coil of sufficiently high density results. The direction in which one finds the first carbon atom of the side chains for amino acids, other than proline or hydroxy proline, is indicated on the plan of Fig. 1 by the symbol R. A single turn of the α-helix, in which van der Waals atomic radii are included to describe the outer limit of the plan, is shown in Fig. 2. For purposes of discussion, this plan is distorted into a circle of radius equal to 3.4 A. To this circle, one must attach the side chains. An analysis of a variety of proteins, undertaken a few years ago,[2] showed that the average side-chain extension was 5.1 A. Thus, the distance from the helix center to the end of the average extended side chain would be about 8.5 A.

At this stage, certain properties of the side chains should be examined since both internal structure and protein interactions are strongly dependent on these properties. Figure 3, taken from a review by Low,[3] shows the variations in size and configuration of the amino acids themselves. The side-chain configurations may be visualized by subtracting from each approximately the equivalent of glycine. The nonpolar amino acids have been grouped in the upper left, the upper center are those containing hydroxyl groups, while acidic and basic side-chain groups are in the upper and the lower right, respectively. Table I gives lengths and volumes for twenty common amino-acid side chains. An examination of the second column, which gives extended length, reveals that side chains vary from 1.5 A for glycine to 8.8 A for arginine. The volumes vary from 5.1 A^3 for glycine to 175.5 A^3 for tryptophan. Figure 4 shows that, in general, there is a closely linear relationship between side-chain length and volume except for the larger nonpolar side chains of valine, leucine, isoleucine, phenylalanine, tyrosine, and tryptophan. All of these are significantly *shorter* for the volumes occupied. I feel that the reason for this situation is to prevent, under ordinary circumstances, interactions of the larger nonpolar side chains. Thus, an alteration in structure generally is necessary to allow these side chains to come into contact.

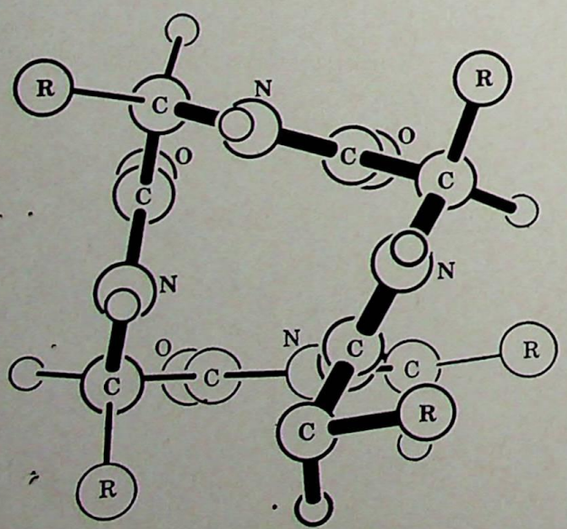


FIG. 1. Plan of the 3.7-residue α-helix [from L. Pauling, R. B. Corey, and H. R. Branson, Proc. Natl. Acad. Sci. U. S. 37, 205 (1951)].

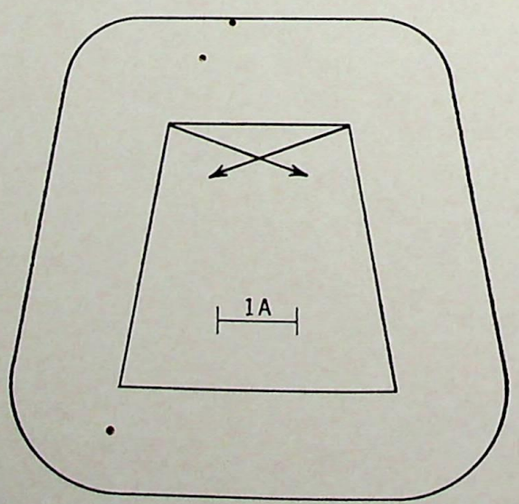


FIG. 2. Plan view of one turn of the α-helix including van der Waals radii [from D. F. Waugh, Advances in Protein Chem. 9, 326 (1954)].

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

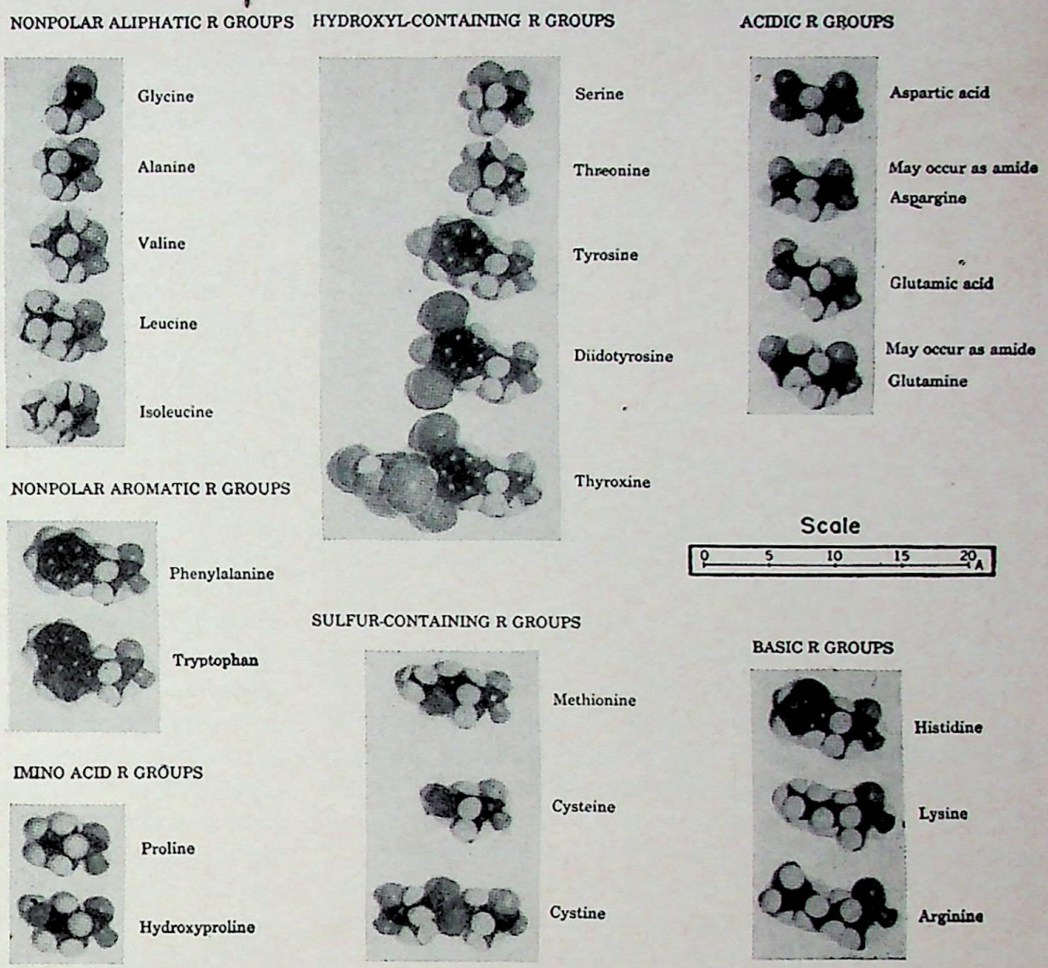


FIG. 3. Packing models of amino acids as dipolar ions [from B. W. Low in *The Proteins*, H. Neurath and K. Bailey, editors (Academic Press, Inc., New York, 1953), Vol. I, p. 235].

The coiled main chain is viewed as a tube, flat-sided in the case of the α-helix, having protuberances of varying shape and volume. The shapes and spatial positions of the ring-containing side chains are fixed except for the freedom permitted by rotation around the bonds of the CH_2 group which, except for proline and hydroxy proline, joins the ring structure to the main chain. The positions of the protuberances will depend on the amino-acid sequence and the precise nature of the main-chain coil through that position.

In many, if not most proteins, the polypeptide chain occurs in segments or folds, and in the native molecule these segments or folds are approximated and bonded laterally. Certain of the lateral bonds may be covalent bonds, the primary focus of attention in this respect being the disulfide bonds of cystine. Others will be ionic interactions and secondary valence interactions such as hydrogen bonds and short-range van der Waals forces. Comments about the interactions of chain segments in producing the protein molecule are, I believe, applicable with some modification to the interactions between protein molecules.

Figure 5 gives a diagrammatic representation of a cross section through a four-chain molecule. This molecule is constructed on the assumption that, at least for short distances, the chains are in the form of straight columns and that the center-to-center distance between chains is about 10 A, a value heretofore generally ac-

TABLE I. Characteristics of amino-acid side chains.

Amino acid	Length max, A	Volume, A³	Amino acid	Length max, A	Volume, A³
Aspartic acid	5.0	58.4	Glycine	1.5	5.1
Asp. (amide)	5.1	65.4	Alanine	2.8	32.2
Glutamic acid	6.3	85.5	Serine	3.8	36.0
Glu. (amide)	6.4	92.5	Threonine	4.0	63.1
Arginine	8.8	125.7	Methionine	6.9	112.1
Histidine	6.5	89.0	Valine	4.0	86.3
Lysine	7.7	121.0	Leucine	5.3	113.4
Cystine ½	2.9	52.8	Isoleucine	5.3	113.4
Cysteine	4.3	57.9	Phenylalanine	6.9	136.6
Tyrosine	7.7	138.8	Tryptophan	8.1	175.5

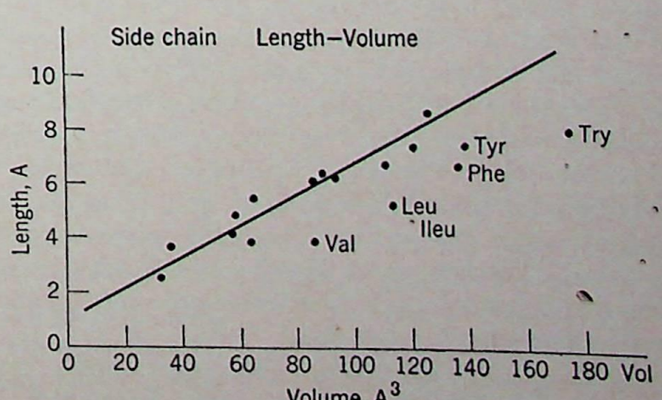


FIG. 4. Side-chain length *vs* volume.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

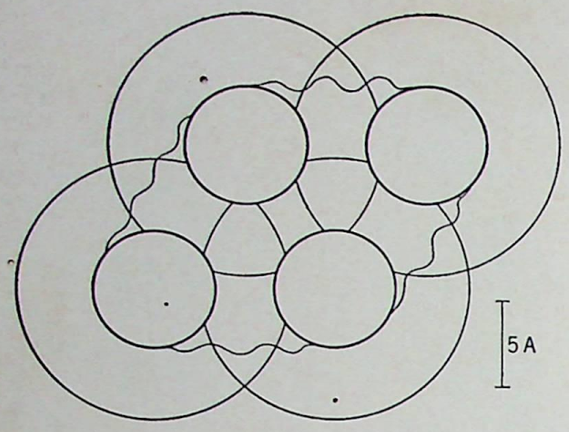


FIG. 5. Schematic representation of the interaction of four helices to produce an internal volume (enclosed by wavy lines) in which side chains are closely packed. Heavy circles are main-chain domains and light circles give maximum extension of average side chain [from D. F. Waugh, Advances in Protein Chem. 9, 326 (1954)].

cepted from crystallographic data. It is probable that the axes of the helices are not linear and that the center-to-center distance is not always 10 A, but the diagrammatic representation is examined for what it suggests as to the way in which the side chains are packed in the protein molecule.

Figure 5 shows heavy circles which are the main-chain domain, and light circles which give the average maximum side-chain extension calculated from the amino-acid composition of insulin. The wavy lines demarcate an internal volume, where side chains are closely packed, from surface regions where side chains are brought into contact with other side chains usually through interactions with other molecules. The side-chain volume enclosed within the wavy lines is expected to be crowded, and thus any regularity in the positioning of the main chain would necessitate a careful selection of the size, shape, and interaction properties of the side chains present on each of the four contributing helices. The precision of this selection is appreciated when an examination is made of the amount of unoccupied space in the protein molecule. By unoccupied space is meant a space which is too small to accept a water molecule or a larger space which is not accessible to the solvent. The maximum unoccupied space may be assessed as follows.

First, from a comparison of the measured specific volume of a protein with that calculated by Traubes' rules from the amino-acid composition, McMeekin and Marshall[4] find that the two values are in excellent agreement for a variety of proteins, as shown in Table II. As Edsall[5] has pointed out, however, the measured value should be about 3.5% lower than the calculated value because of the electrostriction of water around charged groups. The difference one may attribute to unoccupied space.

Second, Linderstrøm-Lang[6,7] has shown for several proteins that enzymatic cleavage of the first few peptide bonds leads to a volume decrement in excess of that predicted from electrostriction of water around the new charged groups produced. The excess volume decrement is about 3% of the molar volume.

The internal volume of the four-chain molecule shown in Fig. 5 is about one-third of the total volume; thus, if all of the unoccupied space occurs in the internal volume, about 10% of the latter will be unoccupied.

A rough calculation, based on the average side-chain volume and the frequency with which side chains are expected to occur in the internal volume, also suggests that roughly 10% of the internal volume will be unoccupied.[2]

The quantity of unoccupied space calculated in the foregoing is expected to be a maximum value. For example, disruption of the protein structure may alter the structure of the surrounding solvent so that a decrease in volume occurs. If such is the case, the unoccupied space in the internal volume would be less than the 10% already suggested, and as a result the side chains in the internal volume would necessarily be more closely packed. Linderstrøm-Lang[6] has pointed out that the protein molecule has some forced structure which makes it occupy more space than the unfolded elements. This forced structure is, I feel, the result of a compromise between uniformity in main-chain configuration and perfection in side-chain packing.

Here are a few additional remarks about the internal volume. In native proteins, it is the rule—rather than the exception—to find that a portion of groups such as sulfhydryl, disulfide, phenol, imidazol, etc., appear to be altered strikingly either in their physical or in their chemical properties. These have been referred to as hidden groups.

Interest in hidden groups has generally been associated with interest in the more general problems of protein denaturation (see the following), for denaturation, reversible or otherwise, gives at least a portion of the hidden groups a freedom to interact in more customary fashion.[8–10] Thus, such groups in native proteins may not react chemically as would be expected from

TABLE II. Specific volumes of proteins [from T. L. McMeekin and K. Marshall, Science 116, 142 (1952)].

Protein	Observed volume cc/g	Calculated volume cc/g
Ribonuclease	0.709	0.703
Lysozyme	0.722	0.717
Fibrinogen (human)	0.725	0.723
α-casein	0.728	0.725
Chymotrypsinogen	0.73	0.734
Serum albumin (bovine)	0.734	0.734
Insulin	0.735	0.724
β-casein	0.741	0.743
Ovalbumin	0.745	0.738
Hemoglobin (horse)	0.749	0.741
β-lactoglobulin	0.751	0.746
Edestin	0.744	0.719
Botulinus toxin	0.75	0.736
Gelatin	0.682	0.707

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

TABLE III. Average side-chain nonpolarity and nonpolar side-chain frequencies.

Protein	Av side chain Equiv. CH_2	Nonpolar class Equiv. CH_2	Nonpolar class Occurrence frequency
Collagen	1.33	4.06	0.21
Rabbit tropomyosin	2.15	4.32	0.25
Ribonuclease	2.16	4.33	0.26
Rabbit myosin	2.28	4.41	0.28
Calf-liver histone	2.29	4.13	0.31
Chymotrypsinogen	2.29	4.32	0.33
Human fibrinogen	2.31	4.53	0.32
Ovalbumin	2.33	4.32	0.36
Bovine serum albumin	2.37	4.74	0.32
Edestin	2.37	4.31	0.33
Horse myoglobin	2.39	4.32	0.36
β-lactoglobin	2.41	4.24	0.37
FSH	2.44	4.27	0.39
Horse hemoglobin	2.46	4.19	0.43
Rat-sarcoma histone	2.47	4.08	0.39
Ox insulin	2.52	4.27	0.45
Human γ-globulin	2.54	4.32	0.40
Ox-growth hormone	2.57	4.48	0.39
Calf-thymus histone	2.59	4.21	0.40
α-casein	2.59	4.38	0.40
Human serum albumin	2.61	4.33	0.40
β-casein	2.77	4.08	0.49
Average	2.36	4.30	0.35

the behavior of the same groups in smaller molecules, the ultraviolet absorption spectra may differ from those determined with less complicated molecules, and the acceptance of protons by basic groups may be retarded. Most of these effects must be associated with that portion of the molecular structure capable of bringing side chains into juxtaposition and of shielding either through secondary valence interactions or through steric hindrance. In a four-chain molecule, approximately half of the side chains will be involved in interactions within the internal volume, the other half will, to a large extent, define the configuration and surface properties of the molecule.

Examination of a series of proteins reveals that the amino acids are distributed as follows: 0.2 to 0.5 can carry charges (are, therefore, in the groups arginine, histidine, and lysine which may carry positive charge; aspartic acid, glutamic acid, cysteine, and tyrosine which may carry negative charge). The smaller amino acids such as glycine, alanine, threonine, and methionine make up 0.13 to 0.5 of the side chains, and the larger nonpolar amino acids of valine, leucine, isoleucine, proline, phenylalanine, tryptyophane, and tyrosine contribute 0.21 to 0.47 of the side chains. There appear to be no striking correlations between the average occurrences of various types of side chains, other than that as the frequency of one type increases, the frequencies of the other types decrease. However, the average side-chain volume may be kept relatively constant. This is suggested by the fact that the average side-chain volumes for a group of five proteins varied over a range of only 6% (Table V in reference 2).

The interactions of proteins with each other must, to a considerable extent, depend upon the characteristics of side chains although segments or regions of exposed main-chain groups might also be involved. A brief summary of the interaction characteristics of charged groups, hydrogen bonds, and van der Waals forces has been given.[2] Of particular interest are the following special possibilities. Kirkwood and Shumaker[11] show that the mobile protons on a group of particles, at *p*H values close to the isoelectric point and at low ionic strength, will effect charge patterns which will produce a net attractive force. At ionic strengths in the physiological range (~0.15) and at *p*H values removed by more than a *p*H unit from the isoelectric point, interactions of charges are expected to give rise to repulsion and the hydration of charged groups to a barrier which will modify the interactions of those groups dependent upon close approach, such as hydrogen-bond-forming groups and nonpolar groups.

The formation of an interprotein or intraprotein hydrogen bond is usually an exhange reaction involving the interacting groups and water. As shown by Pauling and Pressman[12] and by Schellman,[13] the resulting interaction energy is about 1 kcal/mole; thus, several such bonds, acting alone, would be required to give a stable protein-protein interaction.

The interactions of a few of the larger nonpolar groups with the concomitant formation of new hydrogen bonds in the released water, would suffice to form a stable intermolecular linkage. This conclusion stems from the fact that the association of hydrocarbon chains in an aqueous environment liberates about the same energy as the condensation of the hydrocarbon from the vapor state—namely, about 1.2 kcal/mole of CH_2 groups. Thus, the interaction, under the proper conditions, of a single CH_2 group may be equivalent to the formation of an interprotein hydrogen bond.[2] Important also, from the standpoint of internal structure and stability as well as of interaction, is the insensitivity of the nonpolar interaction itself to the *p*H or ionic strength of an aqueous environment.

The average, large, nonpolar side chain is the equivalent of about 4.2 CH_2 groups. Table III, which lists the frequency of occurrence of large nonpolar side chains for a group of proteins, shows an average value of 0.35, the minimum for corpuscular proteins being 0.25 for highly soluble tropomyosin and 0.49 for β-casein. It is clear that the interaction of a small fraction of the total number of the larger nonpolar side chains (e.g., 5 to 10 side chains) would be sufficient to produce an interaction product which would be stable over a wide range in *p*H near the isoelectric point. Some proteins, like serum

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

albumin, are soluble at their isoelectric points and most are soluble to an appreciable extent at *p*H values one or two units away from the isoelectric point. It is thus apparent that the native-protein structure (or family of structures) positions most of the groups capable of attraction so that they are shielded; in other words, so that the interactions of attractive groups are opposed by energy-requiring processes such as (a) the approximation of groups carrying like charges, (b) the removal of water of hydration from dipolar or charged groups, (c) the breaking of hydrogen bonds between water and protein groups, and (d) a distortion of the helix (secondary) structure or side-chain interactions (tertiary structure[7]), etc. In this connection, it is significant that the nonpolar side chains are shorter per unit volume than other side chains and that the charged side chains, particularly those carrying positive charges, are somewhat longer. The addition of water of hydration to charged side chains would cause all of these to project beyond the values given in Table I.

The native protein is extracted by using as mild as possible a set of physical and chemical conditions, experience being required to define the sets of conditions which can be considered mild for a given protein. The protein is characterized, so far as is possible, also using mild conditions, the end result being a description of the protein in terms of its physical, chemical, and biological properties. If severe conditions of treatment are chosen (such as *p*H, temperature, the presence of urea, guanidine, salicylate, etc.), the properties of the native protein change. One of the most striking alterations is in solubility, and initially the term denaturation was based on the fact that almost all proteins become insoluble after a variety of severe treatments. In general, denaturation should not involve hydrolysis of protein covalent bonds. The importance of and interest in denaturation are indicated by the sequence of reviews which have treated this subject.[8–10,14,15]

It is instructive to examine a typical experiment, involving denaturation by alkali, in which case experience dictates an appropriate choice of *p*H, ionic strength, and temperature.

Figure 6, in the left-hand column, indicates that the native protein under ordinary conditions has a low net charge and that establishing denaturing conditions first produces a native protein of high net charge (upper left). With time, the properties of the protein population alter according to first-order kinetics to give reversibly denatured molecules also of high net charge. Under the conditions of denaturation, the protein usually remains in solution, aggregation and precipitation being prevented by the electrostatic energy barrier to close approach. That alterations in structure have occurred is demonstrated by taking an aliquot of the protein solution under denaturing conditions and, by diluting with an appropriate buffer, rapidly returning to a *p*H near the isoelectric point where the native protein is soluble. The fraction of the protein which has been denatured will precipitate. Once a precipitate is allowed to form, the precipitate is insoluble over a broad *p*H range including alkaline *p*H values where the denatured protein remains in solution on careful downward adjustment of the denaturing *p*H. Not only is this general, but if the precipitate is tested with respect to time, the limits of *p*H over which insolubility is observed are broadened, either indicating an increasing number of participating bonding groups or indicating that slow rearrangements take place which make existing group participations more effective. The insolubility of denatured proteins is attributed by many primarily to nonpolar interactions, for, since the driving energy is actually the formation of new water-hydrogen bonds, these are the group interactions which are particularly insensitive to changes in *p*H and ionic strength. Clearly, the conformations of the denatured molecules do not preserve the balance between attractive group interactions and energy-requiring processes discussed in the foregoing in connection with the native protein, alterations in the structures of the latter generally producing local patches of attractive groups which are no longer shielded.

That the denatured protein has a structure different from that of the native protein may also be shown under the denaturing conditions—for example, by changes in ultraviolet absorption, optical rotation, chemical reactivity of particular groups, and by applying techniques designed to examine size, shape, and hydration. Occasionally, determinations of biological activity may also be made under the conditions of denaturation.

If the protein is denatured as described—i.e., when the molecules are prevented from interacting by hydration and charge repulsion—a slow or stepwise return to conditions where the protein is native may lead to a reversal of denaturation, that is, to a return of solubility, biological activity, normal optical activity, etc.

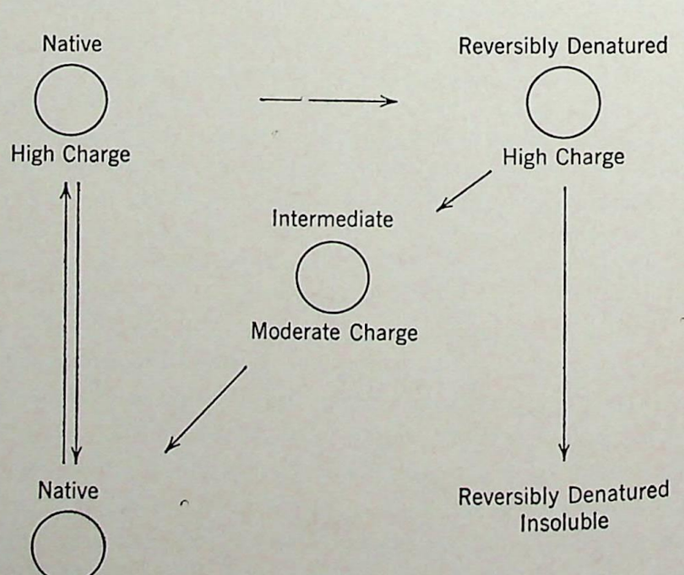


FIG. 6. Scheme showing the manipulations involved in demonstrating reversible denaturation by alkali.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

A situation frequently encountered is illustrated by the central diagram of Fig. 6. An aliquot of the solution under denaturing conditions (reversibly denatured, high charge) is brought to a critical *p*H intermediate between the denaturing *p*H and a lower *p*H where the denatured protein will precipitate. At this intermediate *p*H, the molecules are somewhat expanded, the behavior of the system suggesting that a compromise has been reached: progressively below this narrowly defined *p*H, activation energy barriers to recovery increase (and thus effectively freeze the conformations of the denatured states), while above this *p*H, conformation changes corresponding to denaturation are increasingly favored. Clearly, the set of conformations which are established with time at the intermediate *p*H will, on reducing the *p*H further, tend to produce conformations corresponding to the native state. For the protein fibrinogen, denaturation is carried out at *p*H 12.4, ionic strength 0.15, and 0°C. Denaturation by a slow reaction is largely complete in 60 min. The critical intermediate *p*H is 10.8 and a treatment time of 16 hr at 0°C is necessary to effect a recovery of native properties. As is so often the case, however, the recovered protein resembles native protein in many gross aspects—for example, clottability after treatment with thrombin—but differs in detail—for example, in optical activity or solubility at *p*H values close to the isoelectric point. After reversal of denaturation, few proteins have been shown to be identical in all respects with the original native protein.

For many proteins, conditions may be chosen so that an equilibrium exists between the native and denatured states. For example, the enzyme trypsin[14,16,17] at a *p*H near *p*H 11 and 50°C shows a distribution between native and denatured states, for, on adding salt and rapidly altering the *p*H to *p*H 7, a fraction of the protein precipitates while a fraction remains soluble and exhibits enzymatic activity. The distribution is sensitive to temperature, the properties of the system suggesting a true equilibrium, a fact which allows a determination of the free energy, enthalpy, and entropy of the reaction. The data[18] for soybean trypsin inhibitor are shown in Table IV (see also reference 14). It is quite clear that a relatively small change in free energy is the result of a balance between large positive changes in enthalpy and entropy. Essentially the same result is obtained when other proteins showing temperature-sensitive denaturation equilibria are examined (trypsin, chymotrypsinogen, pepsin, luciferase; see reference 14). The balance between large entropy and enthalpy changes (of either sign) to yield small free-energy changes is quite apparent also in the rate-defining parameters of the forward- and backward-activation reactions. So far, denaturation reactions have been too complicated to be analyzed in precise structure terms. The present concepts involve mechanisms in which there are:

(1) Alterations in proton association and attending water of hydration.

(2) Expansion (contraction) effects with concomitant (a) increasing (decreasing) entropies, and (b) increasing (decreasing) heats.

(3) Alterations in tertiary (side-chain) and secondary (main-chain coil) structure leading to an altered availability of surface side chains.

In the foregoing, reference has been made to the fact that interaction generally stabilizes the denatured state, the precipitate of denatured protein becoming more insoluble with time. Precipitation, however, does not necessarily lead to irreversible changes. R. H. Hartley and I have denatured fibrinogen by alkali and precipitated the denatured protein by rapid alteration in *p*H. The precipitate was recovered by centrifugation and redissolved in alkali at *p*H 12.4. It was then brought to *p*H 10.8 and allowed to stand for several hours at 0°C after which the *p*H was returned to *p*H 7. The majority of the protein was then found to be soluble and clottable with thrombin, although it differed somewhat from native fibrinogen in its precipitation at the isoelectric point.

So far, those interactions have been considered which, I feel, are generally accepted as involving, relatively, a disorganization of structure and thus the distribution of side chains. A well-organized site of interaction is clearly indicated in interactions between antibody molecules and hapten groups as antigens, a subject which has been extensively studied by Pauling, Campbell, Pressman, and their associates. For example, a protein such as serum albumin is coupled with the diazonium salt of a molecule to give substitutions such as azobenzenearsonates, (hydroxyphenyl azo) benzoates, azobenzoates, etc. Antibodies, found in the γ-globulin fraction of the plasma proteins, appear when these modified proteins are injected into animals. The antigens are found to combine relatively specifically with the coupled azo group.[19]

A variety of studies has been made of the combination of hapten with antibody and of the way in which various types of molecules inhibit the formation of an antigen-antibody precipitate.[20,21] The inhibiting action is most effective when the structure of the hapten is duplicated and becomes less effective as the inhibiting molecule differs in the positions of the groups attached to the benzene ring (the substituent groups), in the sizes of these groups (assuming similar chemical affinity), and in chemical affinity. The conclusion has been reached that the main-chain coils of the antibody molecule are

TABLE IV. Soybean trypsin inhibitor *p*H = 3 T = 40°C [from M. Kunitz, J. Gen. Phys. 32, 241 (1948)].

	Thermodynamic	Rate parameters: Forward reaction	Rate parameters: Back reaction
ΔF, kcal/mole	1	25.4	24.4
ΔH, kcal/mole	57.3	55.3	− 1.9
ΔS, cal/mole/deg	180.0	99.5	−85

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

arranged so that they enclose a cavity or slot which accommodates the hapten group and follows the contours of the latter within a few tenths of an Ångström unit. Interaction has been shown to involve whatever forces are permissible from the properties of the hapten: nonpolar forces if the hapten has no polar or dipolar groups, hydrogen bonds where donor (acceptor groups) are present on the hapten, and charge interactions when such can be made.

What data are available suggest that the free-energy change accompanying a reaction such as

$$\begin{aligned} &Ag + AgAb \leftrightarrow Ag_2Ab \\ &Ab\ (ppt) \leftrightarrow Ab\ (sol.) \end{aligned}$$

is small and about 5 to 10 kcal/mole. There are uncertainties concerning changes in entropy and heat. Further considerations are given to antigen-antibody interactions in a later article by Kauzmann (p. 549).

The combination of a hapten-antigen and an antibody can be considered a "point" combination, unless the antibody is directed against the protein surface surrounding the hapten group. In other words, the interacting molecules may have some freedom of rotation around a single axis going through the locus of combination.

Interactions having a specificity which leads to the formation of complexes whose molecular subunits are held in particular positions and orientations with respect to each other are known to occur and some to have particular biological significance. Among these one would place at once the association reactions of insulin[22–24] and α-chymotrypsin,[25] the remarkable series of interactions leading to the high molecular-weight hemocyanins,[26,27] and the interactions of the caseins which lead to the spontaneous formation of stable colloidal micelles.[28,29] The interactions just mentioned lead to symmetrical complexes of limited size, the last two presumably so that they may contribute a relatively small increment in viscosity to the fluid media in which they are transported.

Other interactions lead essentially to extended structures of unlimited size, thus to fibrils. They are commonly referred to as globule-fibril (GF) transformations. Examples of proteins in this group are insulin,[30] tropomyosin,[31] actin,[32] fibrin,[33] and collagen, the latter being treated extensively by F. O. Schmitt (p. 349). These materials are arranged roughly in an order which reflects the increasing difficulty experienced in obtaining the globular (soluble) form of the protein. In addition, tissues such as collagen may contain nonprotein molecules which aid, somewhere in the hierarchy of aggregation, in producing the tissue structure: the tandon or fascia, etc.

The interactions of insulin serve as a specific example which illustrate a more general experience. As described by Oncley (p. 30), the insulin monomer is a unit of $M = 5733$ crosslinked covalently by two disulfide linkages. The maximum, net positive charge at low pH values is 4 units/monomer, a value which should be obtained at $p\text{H} \sim 1$, and can be obtained since insulin is stable in acid solution. The maximum, net negative charge would be 8 charges/monomer, although the instability of insulin in alkali and the high pH at which the two arginine side chains lose their protons would reduce this value to 14 for most experimental conditions. The association reactions of insulin have been carried out at pH 2 where each monomer carries an average, net positive charge of 3.5 units. The monomers of $M = 5733$ form dimers of $M = 11\,466$ so readily that the 5733 monomer is observed only in dilute solutions, after chemical modification, or in solvents such as acetic acid and pyridine (see Yphantis and Waugh[34] for references). Thus, the dimer of $M \sim 12\,000$ is usually considered the interacting unit. These will associate readily in pairs or in groups of three.[22–24] Summarizing statements concerning reversible association would be (1) the electrostatic work involved in bringing molecules together is balanced against short-range attractive forces (hydrogen bonds, nonpolar interactions); (2) at $p\text{H} = 2$, for pairing, $\Delta F_2 = -4.93$ kcal/mole, $\Delta H_2 = -8.1$ kcal/mole, and $\Delta S_2 = -12.1$ cal/mole/deg; and (3) the entropy decrease of -12 eu is far less than the 122 eu which would be expected. Doty and Myers[23] point out that the freeing of water molecules from an association with polar groups would provide the requisite positive entropy change and suggest that 24 water molecules would be sufficient.

Insulin is also one of the smaller proteins which exhibits the GF transformation. It is chosen for discussion here since its kinetic and mechanistic complications may well foreshadow similar complications in other systems, just as the GF transformation with insulin was one of the first to be described. A recent summary has been given.[30]

The insulin fibril forms under conditions where the dimer is the prevalent form, i.e., at pH 2. Heating at 80 to 100°C causes a spontaneous transformation into a population of fibrils, the most numerous, and largest, of which are about 200 Å in diameter and many thousands of Ångströms long. The reaction goes essentially to completion but is reversible in the sense that, under alkaline conditions, the association product, the fibril, disaggregates to yield insulin. Recovery of insulin after disaggregation is evidence that the insulin molecule does not undergo extensive unfolding in the process of forming fibrils. Stronger evidence, and evidence which gives a clue to the mechanism of fibril formation, comes from seeding experiments. In these, preformed fibrils or fibril segments are seeded into insulin solutions at pH 2.0. While such solutions alone are stable for long periods of time at temperatures of 20°C or below, the seeded fibrils or segments recruit insulin from solution and grow in the process. Structurally, the new portions of the fibrils obtained after fibril growth at lower temperatures appear to be identical with those which are formed at 80 to 100°C.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

When a solution of insulin is heated and the transformation of insulin into fibrils is plotted as a first-order reaction (Fig. 7), the resulting curves typically have a lag period which is followed by an essentially linear rise. The extent of the lag period is determined markedly by the initial insulin concentration, as is the slope of the near-linear portion of the curve.

A comparison of the reaction kinetics observed (a) with the growth of seeded fibrils and (b) in the absence of seeding suggests that the fibril is first initiated by a nucleation reaction which involves the cooperative effects of p-interacting units (dimers) according to Eq. (1), where the differential is the rate of nucleation, k_1 is a rate constant, and C is the insulin

$$dn/dt = k_1 C^p \qquad (1)$$

concentration. The value of p is near 3. After initiation, when the fibril has achieved a reasonable size, the fibril grows as a function of its surface area and the free

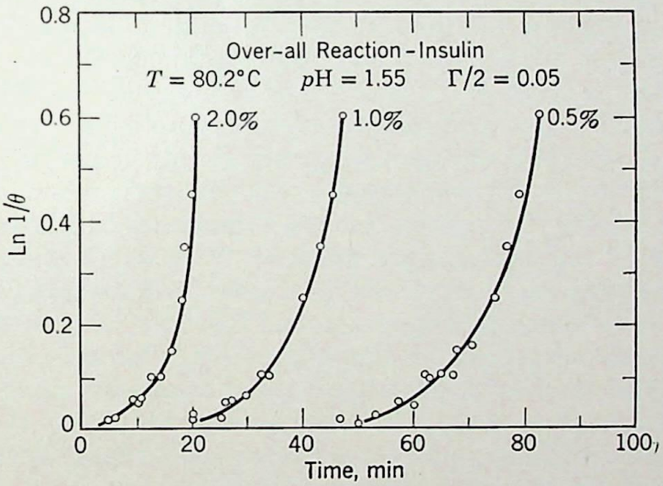


FIG. 7. Kinetics of the transformation of insulin into insulin fibrils. The percent protein has been indicated for each curve. θ represents the fraction of insulin remaining at time t [from D. F. Waugh, J. Cellular Comp. Physiol. 49, Suppl. 1, 145 (1957)].

insulin concentration [Eq. (2)]. The cooperative effect established during nucleation is perpetuated, for the surfaces (particularly the ends) of the fibril present to the entering interaction unit the correct cooperative configuration necessary for bonding.

$$-(dc/dl) = k_2 \text{ area } C. \qquad (2)$$

One particularly interesting consequence of this type of mechanism is that the majority of the fibrils are initiated during the lag period and those which are initiated during the first few minutes of the lag period dominate the reaction in the sense that they are responsible for removing the majority of the insulin. Thus, as shown in Fig. 8 which shows the numbers of fibrils in successive groups plotted against the radius of each group, the fibril population at the end of the reaction appears to be relatively homogeneous.

It is clear that, at least over a portion of its development, the fibril axial ratio increases with fibril mass.

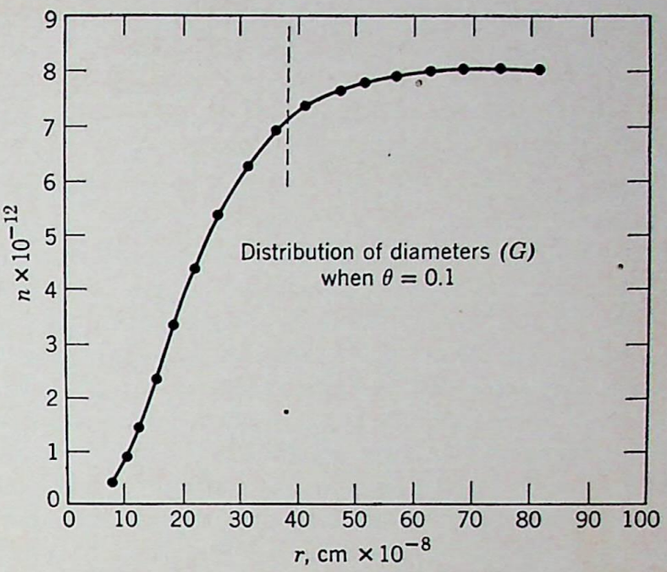


FIG. 8. Distribution of insulin-fibril diameters (G) near the end of a reaction involving 2% insulin when $\theta = 0.1$ [from D. F. Waugh, J. Cellular Comp. Physiol. 49, Suppl. 1, 145 (1957)].

That an asymmetric aggregate should form when charged molecules link, without specification of surface structure, has been pointed out by Rees.[35] This is owing to the fact that the electrostatic-potential barrier is lowest at the ends of the dimer or other asymmetric unit. However, the cooperative effect itself will exert a strong directional influence. This is illustrated in Fig. 9 which is a possible arrangement for the insulin-fibril nucleus. The stable structure is formed when any fourth soluble unit is added to the correct previous group of three. Thereafter, units are expected to add most frequently in a manner which perpetuates this stable

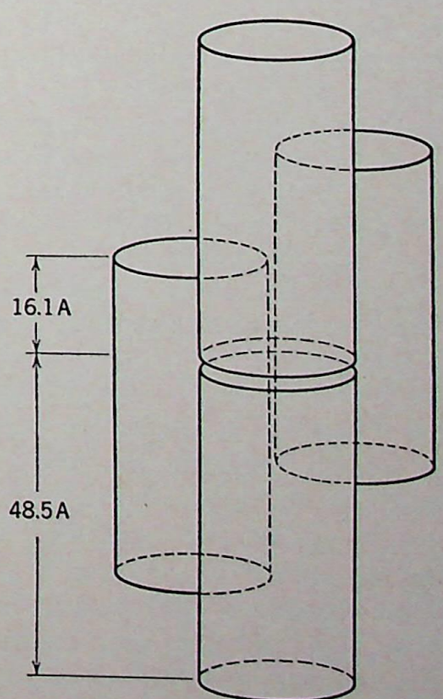


FIG. 9. Diagram of a fibril nucleus illustrating a cooperative interaction [from D. F. Waugh, J. Cellular Comp. Physiol. 49, Suppl. 1, 145 (1957)].

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

structure—consequently, the aggregate elongates in the direction of the cylinder axis of Fig. 9.

A variety of chemically modified insulins will also form insulin fibrils[36]: esterified and acetylated insulins, and those having nonpolar or polar groups added through coupling of insulin with substituted benzenediazonium chlorides. Several different solvents also were used including concentrated urea, organic acids, and ethanol. The results of such studies led to the conclusion that the associations of nonpolar groups were primarily responsible for linking molecules. During the course of these investigations, it was observed also that the introduction of essentially nonpolar groups, for example by reacting benzene- or tolyl-diazonium chloride with the protein, led to insolubility. As the number of groups is increased, the progression is towards larger and more-stable polymers until the insulin itself becomes insoluble. A series of fibrils with increasing stability to disaggregation by alkali is formed also from these insulins.

We have been studying another system in which there occur both specificity of interaction and cooperative effects, the end result being a complex in which secondary valence and charge interactions arrive at a remarkable mutual satisfaction. Reference is made to the interactions of α_s- and κ-caseins.[28,29] Here, α_s-casein is a phosphoprotein of $M = 23\,000$ and has a phosphorus content of 1%. Physically, this protein appears to be a single coil 210 A long and 16 A in diameter. In the absence of calcium, the protein is a highly soluble polymer. The polymer becomes quite insoluble in the presence of calcium. κ-casein is also apparently a single coil about 150 A long, which contains sulfur but little phosphorus. It forms soluble condition-insensitive polymers of 13.5*S* in the presence or absence of calcium. When mixtures are made containing molecular ratios of $3\alpha_s$ to 1κ, the original polymers disappear and a stoichiometric α_s-κ-casein complex forms. This complex can form in the absence of calcium and under conditions where both of the monomers carry a high, net negative charge. One must assume that here, as in other cases, secondary valence forces are responsible for the association. In this connection, an examination of the amino-acid compositions of all of the caseins reveals a high-occurence frequency for the larger nonpolar amino acids mentioned at the start.

The α_s-κ-complex, although it exists at all temperatures between 0 and 37°C, is unstable at the lower temperatures. This is apparent from the results obtained on adding calcium (to give 0.05 M). At the lower temperature, a Ca-α_s-caseinate precipitate forms immediately, leaving κ-casein in solution. At the higher temperature, the complex is stabilized and will now engage in further aggregation to form stable colloidal particles, namely, micelles typical of milk. Once formed, the Ca-α_s-κ-casein complex is unusually stable to heat, changes in *p*H, etc. We propose for it a structure in which the α_s-casein molecules are oriented axially around a κ-casein molecule. At higher temperatures, secondary valence interactions hold the α_s-casein molecules in positions such that pairs of phosphorus groups attached to adjacent α_s-casein monomers are in juxtaposition and can be linked through the introduction of a calcium ion. The precision with which the phosphate groups are positioned is probably dependent upon the conformations of the interactants, for, at lower temperatures, the instability of the complex suggests that the phosphorus groups cannot be crosslinked by calcium. If each phosphorus is able to accept a calcium ion, stabilization through Ca-phosphate crosslinking would be absent, and, since Ca-α_s-caseinate is relatively insoluble at low temperatures, it would precipitate and thus lead to a dissociation of complexes.

Another association which apparently does not involve covalent crosslinks, but does involve specificity of interaction and units of a higher asymmetry than any mentioned so far, is the collagen fibril. According to Boedtker and Doty,[37] the molecule is about 3000 A long and 13.6 A in diameter (hydrated), and is unusual in that it is made up of three polypeptide chains intercoiled to produce a coiled coil. The molecule contains unusually high percentages of hydroxyproline and glycine, and the occurrence frequency for larger nonpolar amino acids, including proline and hydroxyproline, is low (0.31). The collagen molecule denatures in the unusually low temperature range of 25 to 35°C during which process the molecule falls apart irreversibly into its constituent chains, which then fold into more symmetrical structures. A cooperative effect occurs when collagen molecules are linked together to form protofibrils, filaments, etc. (F. O. Schmitt, p. 349). That the cooperative effect exists is shown by the striking difference between the temperature of molecular denaturation and the temperature of thermal shrinkage, the uniform cross-striations, and the diffraction patterns of collagen, etc. Another result of the cooperative effect in collagen is the remarkably high tensile strength of wet collagen fibers. The latter is about 60 kg mm^2 for intact collagen and may be about 80% of this value for fibers reconstituted from collagen gels. These high tensile strengths, in the absence of intermolecular covalent bonds, must be the result of molecular overlapping which allows a maximum of interaction in the formation of secondary valence attractions.

BIBLIOGRAPHY

[1] L. Pauling, R. B. Corey, and H. R. Branson, Proc. Natl. Acad. Sci. U. S. **37**, 205 (1951).

[2] D. F. Waugh, Advances in Protein Chem. **9**, 326 (1954).

[3] B. W. Low in *The Proteins*, H. Neurath and K. Bailey, editors (Academic Press, Inc., New York, 1953), Vol. I, p. 235.

[4] T. L. McMeekin and K. Marshall, Science **116**, 142 (1952).

[5] J. T. Edsall in *The Proteins*, H. Neurath and K. Bailey, editors (Academic Press, Inc., New York, 1953), Vol. I, p. 549.

[6] K. Linderstrøm-Lang, Cold Spring Harbor Symposia Quant. Biol. **14**, 117 (1950).

[7] K. Linderstrøm-Lang in *Lane Medical Lectures, 1951* (Stanford University Press, Stanford, California, 1952).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

[8] H. Neurath, J. P. Greenstein, F. W. Putnam, and J. O. Erickson, Chem. Revs. **34**, 157 (1944).
[9] M. L. Anson, Advances in Protein Chem. **2**, 361 (1945).
[10] J. Steinhardt and E. M. Zaiser, Advances in Protein Chem. **10**, 151 (1955).
[11] J. G. Kirkwood and J. B. Shumaker, Proc. Natl. Acad. Sci. U. S. **38**, 863 (1952).
[12] L. Pauling and D. Pressman, J. Am. Chem. Soc. **67**, 1003 (1945).
[13] J. A. Schellman, Compt. rend. trav. Lab. Carlsberg. Sér. chim. **29**, 223 (1955). See also, W. F. Harrington and J. A. Schellman, *ibid.* **30**, 21 (1956).
[14] R. Lumry and H. Eyring, J. Phys. Chem. **58**, 110 (1954).
[15] F. W. Putman in *The Proteins*, H. Neurath and K. Bailey, editors (Academic Press, Inc., New York, 1953), Vol. I, p. 893.
[16] A. S. Stern, Ergeb. Enzymforsch. **7**, 1 (1938).
[17] M. L. Anson and A. E. Mirsky, J. Gen. Physiol. **17**, 393 (1934).
[18] M. Kunitz, J. Gen. Physiol. **31**, 241 (1948).
[19] K. Landsteiner, *Specificity of Serological Reactions* (Harvard University Press, Cambridge, 1945), revised edition.
[20] D. H. Campbell and N. Bulman in *Fortschritte der Chemie organischer Naturstoffe*, L. Zechmeister, editor (Springer-Verlag, Wien, 1952), Vol. IX, p. 443.
[21] D. Pressman in *Molecular Structure and Biological Specificity*, L. Pauling and H. Itano, editors (American Institute of Biological Sciences, Publication No. 2, Baltimore, Maryland, 1957), p. 1.
[22] J. L. Oncley, E. Ellenbogen, D. Giflin, and F. R. N. Gurd, J. Phys. Chem. **56**, 85 (1952).
[23] P. Doty and G. E. Myers, Discussions Faraday Soc. **13**, 51 (1953).
[24] R. F. Steiner, Arch. Biochem. Biophys. **44**, 120 (1953).
[25] G. W. Schwert and S. Kaufman, J. Biol. Chem. **190**, 807 (1952).
[26] I. Ericsson-Quensel and T. Svedberg, Biol. Bull. **71**, 498 (1936).
[27] A. Tiselius and F. L. Horsfall, Jr., J. Exptl. Med. **69**, 83 (1939).
[28] D. F. Waugh and P. H. Von Hippel, J. Am. Chem. Soc. **78**, 4576 (1956).
[29] D. F. Waugh, Discussions Faraday Soc. **25**, 186 (1958).
[30] D. F. Waugh, J. Cellular Comp. Physiol. **49**, Suppl. 1, 145 (1957).
[31] T. C. Tsao and K. Bailey, Discussions Faraday Soc. **13**, 145 (1953).
[32] S. V. Perry, Physiol. Rev. **36**, 1 (1956).
[33] H. Scheraga and M. Laskowski, Jr., Advances in Protein Chem. **12**, 1 (1957).
[34] D. A. Yphantis and D. F. Waugh, Biochim. et Biophys. Acta **26**, 218 (1957).
[35] A. L. G. Rees, J. Phys. & Colloid Chem. **55**, 1340 (1951).
[36] D. F. Waugh, D. F. Wilhelmson, S. L. Commerford, and M. L. Sackler, J. Am. Chem. Soc. **75**, 2592 (1953).
[37] H. Boedtker and P. Doty, J. Am. Chem. Soc. **78**, 4267 (1956).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

9

Three-Dimensional Structure of Globular Proteins

JOHN C. KENDREW

Cavendish Laboratory, The University of Cambridge, Cambridge, England

PRECEDING articles (Oncley, p. 30; Rich, p. 50; Waugh, p. 84) have presented protein molecules in various ways. With an eye to the properties they wished to describe, the authors have used conceptual schemes ranging from the straight polypeptide chain to the helical configurations, and finally to the unextended, bundled-up shape of the globular proteins. These different configurational aspects must be studied in different ways. Chemical methods are used to study the sequence of amino-acid residues in the polypeptide chain. Indications of the helical nature of the chain can be obtained, for example, by measuring optical rotations. Finally, the three-dimensional configuration of a globular protein, the detailed twisting and folding of a polypeptide chain to form a bundle of characteristic shape, can be determined only by x-ray diffraction.

Proteins are, of course, much more complicated than the molecules usually studied by the x-ray method. Use of this method began in 1913 with Bragg's determination of the structure of sodium chloride, which has just two atoms in its molecule. The most complicated complete structure so far determined—vitamin B_{12}, studied by D. Hodgkin—has 90 some atoms in its molecule (exclusive of hydrogens). Even the smallest proteins are much larger than vitamin B_{12}. Their great size and complexity would, in fact, soon discourage the crystallographer, if the importance of the problem were not so great; for it is not difficult to see that in such highly complicated molecules a knowledge of the three-dimensional configuration is vitally important for an understanding of physicochemical properties or biological function.

X-rays are scattered by electrons, and the amount of scattering of x-rays at any point in a crystal is proportional to the electron density at that point. A crystal is essentially a three-dimensional periodically repeating structure. The electron density plotted as a function of distance along a crystal axis is, therefore, periodic and can be analyzed into a series of harmonics, just as a musical note can be analyzed into its fundamental and overtone components. To carry the analogy further, a complicated sound can be reconstituted by a so-called Fourier synthesis or putting together of its component waves (Fourier components). In order to synthesize a musical note, one needs certain items of information about each of the harmonics: their frequencies, their amplitudes, and also their phases, that is, how much they are out of step with one another.

In x-ray diffraction, one must generalize this sort of argument into three dimensions. The complete diffraction pattern of a crystalline protein consists of a regular three-dimensional array of spots, any plane of which can be sampled by means of an x-ray photograph (Fig. 1). The great usefulness of the diffraction pattern lies in the fact that, for any crystal, each of the spots corresponds to a separate Fourier component of the periodically repeating electron-density distribution of that crystal. The complete distribution is the sum of all such waves, each having a characteristic amplitude, wavelength and direction relative to the crystal axes, and phase. The length and direction of the wave can be determined directly from the position of the spot on the photograph and the geometry of the x-ray camera, and the amplitude can easily be inferred from the blackness of the spot on the photograph. It is the determination of the remaining parameter, the relative phase of the spot, which presents the biggest difficulties in x-ray crystallography, for no direct physical method of phase determination is available.

In the study of simpler structures, phases are usually found by trial-and-error methods. The crystallographer guesses the atomic positions before he begins the analysis. If his guess is good enough, he can predict the phase relationships and then begin a mathematical refinement process in which he gradually distorts the proposed model until the diffraction pattern predicted from it matches that obtained experimentally. He does this by minimizing certain error functions which express the discrepancies between the observed and predicted pat-

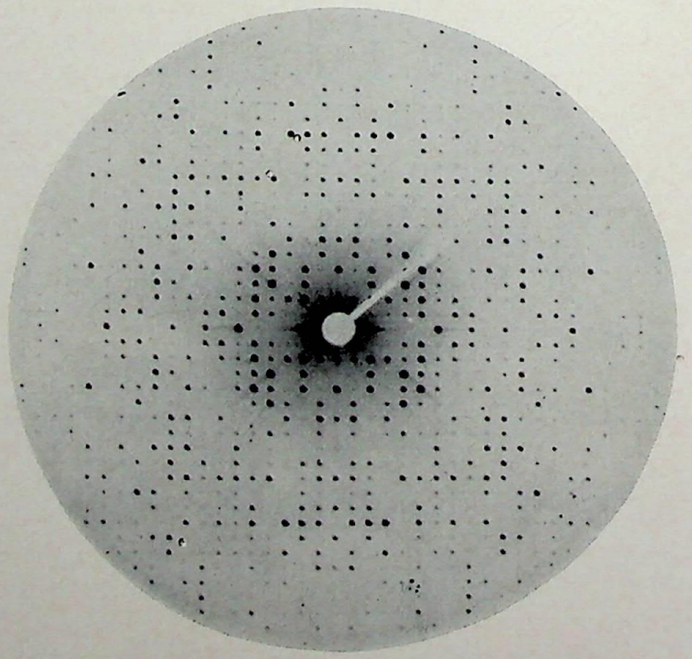


FIG. 1. X-ray photograph of sperm-whale myoglobin crystal.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

terns. In general, the method can yield any one of a number of solutions, and the correct one will be found only if the initial guess was a fairly good one. For complex molecules, such as proteins, the method has been found to be useless, for the number of possible solutions in such cases is so great that there is negligible probability that any initial guess as to the positions of all of the atoms is sufficiently near the truth to be refined into the correct solution.

Fortunately, there are a few tricks that the crystallographer can use for determining phases in more complicated structures. One such trick, the method of isomorphous replacement, is to replace an atom or small group in the molecule by a very heavy atom or group. If the replacement causes an appreciable change in the diffraction pattern; if the positions of the heavy groups in the unit cell are known; and if the replacement causes no distortion of the original crystal structure, it is then possible to determine phases by comparing diffraction patterns before and after replacement.

The isomorphous-replacement method was first used successfully in protein crystallography in 1953, when M. F. Perutz applied it to horse hemoglobin. This protein molecule contains sulfhydryl groups which can react specifically with reagents such as *p*-chloromercuribenzoic acid (PCMB), and Perutz found a definite change in the diffraction pattern when the reaction took place. For the method to work, it is essential that only a small number of heavy groups are attached to each molecule and that they attach at specific sites. When these conditions are met, one can establish the positions of the heavy atoms in the unit cell by straightforward methods. From the weights of the heavy groups and from their positions, one can calculate their vectorial contribution (i.e., both phase and amplitude) to the harmonic wave represented by each of the spots in the diffraction pattern. Corresponding waves for the protein alone, and for protein plus heavy atom, are represented by vectors of known (experimentally determined) amplitudes, but unknown phase angles; however, since the vectorial difference between them (i.e., the contribution of the heavy groups) is completely determined, one can solve for their phase angles geometrically by finding an arrangement in which the two vectors of specified length form a triangle with the difference vector. None of these operations requires any guesswork.

Perutz determined some phase angles for hemoglobin in this way and proceeded to calculate a two-dimensional Fourier synthesis. He confined his attention to a small class of reflections present in the diffraction pattern of hemoglobin (and most other crystals) where the symmetry of the crystal restricts the phase angles to values of either 0 or π radians. When the Fourier synthesis is carried out with just these reflections, a plane projection of the three-dimensional electron-density distribution is obtained. By working in two

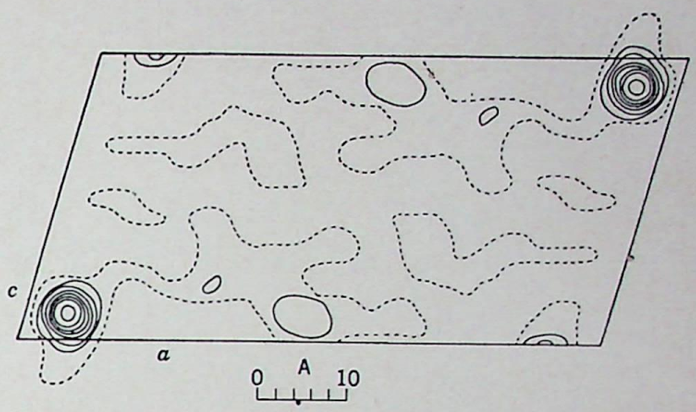


FIG. 2. Difference Fourier projection of the complex of mercury diammine with myoglobin. The unit cell contains two protein molecules; the two peaks indicate that one mercury atom is attached to each molecule.

dimensions instead of three, he achieved a twofold simplification: only a few of the reflections had to be considered, and the phases of these were relatively easy to determine. It is obviously easier to determine a variable with just two possible values than it is to determine one which may have any value.

By putting other heavy atoms in different positions in the molecule, Perutz was able to redetermine the same phases by an independent set of calculations. The results were the same, showing that his projection of hemoglobin was quite certainly correct; unfortunately, however, it proved impossible to interpret in terms of chemical structure. The unit cell is about 40 atoms thick, and in the projection all of these atoms came on top of one another, so that the features of the molecule obscured one another in a hopeless confusion. Obviously, the only way of getting useful information was to extend this analysis into three dimensions.

I now turn to our own studies of myoglobin, which like those of hemoglobin began with a two-dimensional analysis. We chose myoglobin for our study because it is a small protein, having a molecular weight of only 18 000. The molecule consists of 153 amino-acid residues, all arranged, so far as we know, in a single polypeptide chain to which is attached one heme group. This group is common to hemoglobin and myoglobin, and has the property of reversible combination with oxygen. While hemoglobin is used for transporting oxygen in the blood stream, myoglobin is intracellular and its function is to act as a temporary store of oxygen. Though myoglobin is present in all mammals, it is especially abundant in animals such as whales which spend a great amount of time under water and must store oxygen for long periods.

There are several heavy groups which form good isomorphous-replacement compounds with myoglobin. In order to find out whether a heavy group has attached to specific sites in the myoglobin crystal, one takes an x-ray diffraction picture of the protein with heavy group added and another of the protein alone. One then, as it were, subtracts one diffraction pattern from the other

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

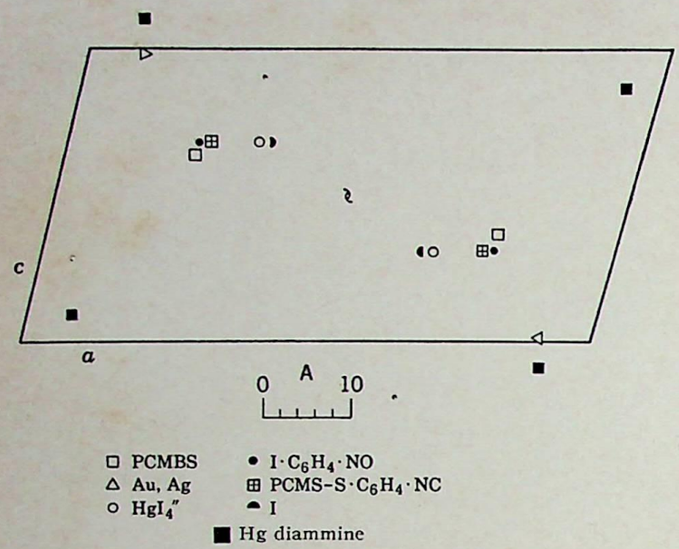


FIG. 3. Sites at which heavy atoms have been attached to the myoglobin molecule. In each case, these are two sites per unit cell, corresponding to one site per molecule.

and carries out a Fourier synthesis in such a way that only the contribution of the heavy group appears in the projected electron-density contour map. If a good isomorphous replacement has been achieved, this two-dimensional difference Fourier synthesis of the unit cell will show a few high peaks and very little else (Fig. 2). A nonspecific absorption of the heavy group or a deformation of the protein would cause the appearance of significant "background" features in the map.

We find, by the foregoing criterion, that myoglobin forms good isomorphous-replacement compounds with mercuri-iodide ion, gold-tetrachloride ion, silver ion, *p*-chloromercuribenzene sulfonate, and mercury diammine. The gold-tetrachloride and silver ions appear to attach at the same site in the molecule. The other groups all go to different sites (Fig. 3). In most of these replacements, we have no idea as to the chemical nature of the attachment, and indeed there are several unexpected features of our results. Thus *p*-chloromercuribenzene sulfonate is known as a reagent for sulfhydryl groups, yet it combines quite specifically to myoglobin which has no free sulfhydryl groups. We do not understand this, nor do we understand why the positively charged silver ion and the negatively charged gold-tetrachloride ion go to the same site in the molecule. There is a great need for some good chemical work on the rationale of combining heavy atoms with proteins. Our approach, however, has simply been to try various reagents and judge our success from the x-ray pictures.

In two-dimensional work, only one replacement compound is really needed, but it is useful to have others because a double check is then possible. For our two-dimensional Fourier projection, we have checked the phases of all of the reflections out to a resolution of about 2 A. We get the same answer with each of four isomorphous compounds, except for a few reflections where the heavy atom contribution is so small that we cannot be sure of its direction. Our two-dimensional Fourier projection, cross-checked four times, was therefore as reliable as Perutz's hemoglobin projection but just as unintelligible (Fig. 4). We were unable to deduce anything about the structure of the myoglobin molecule by looking at it, although now that we have a three-dimensional synthesis we are able to understand in retrospect all of the features of the projection.

We have already noted some of the factors that make it more difficult to work in three dimensions than in two. Besides these difficulties, it turns out that in three dimensions the phases cannot be determined unambiguously if there is only one replacement compound available. On attempting to draw the vector triangle, one finds that there are two possible phase angles consistent with the data for each reflection. However, if the calculations can be done for each of two different replacement compounds, the method yields two pairs of answers of which one from each pair should agree, and these give the correct value of the phase angle. In practice, because of experimental errors and other difficulties, two replacement compounds are hardly sufficient for determining all of the phases. Three compounds or even more are highly desirable.

Before embarking on our three-dimensional synthesis, we had some difficult decisions to make. The amount of work required for such a project goes up very steeply with the degree of resolution sought in the final Fourier synthesis. For myoglobin, we estimated that in order to get a resolution of 6 A one would have to solve the phases of about 400 reflections, while for a resolution of 2 A one would have to treat some 10 000 reflections; to discern the individual atoms in myoglobin, one would need a resolution of at least 1.5 A, and that corresponds to about 20 000 reflections. At this point, the diffraction pattern of a myoglobin crystal begins to fade away altogether. In an ordinary crystal, the diffraction pattern goes on further than this, but a protein crystal is more disordered, and it will probably be touch-and-go as to whether or not we can resolve directly the individual atoms in protein crystals by any method. At the

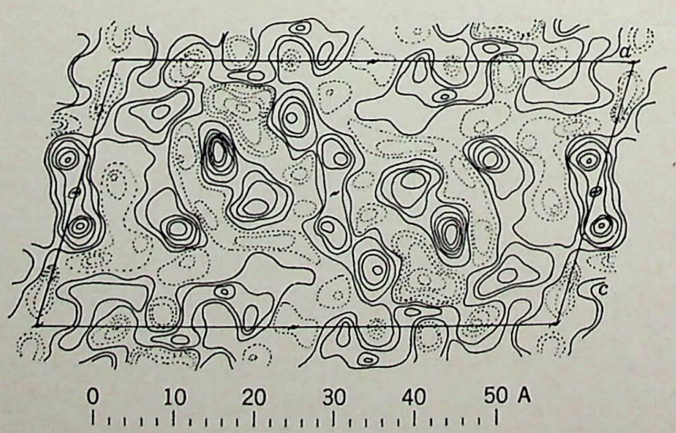


FIG. 4. Two-dimensional Fourier projection of the myoglobin unit cell. The highest peaks in the projection are owing to the iron atoms of the heme group.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

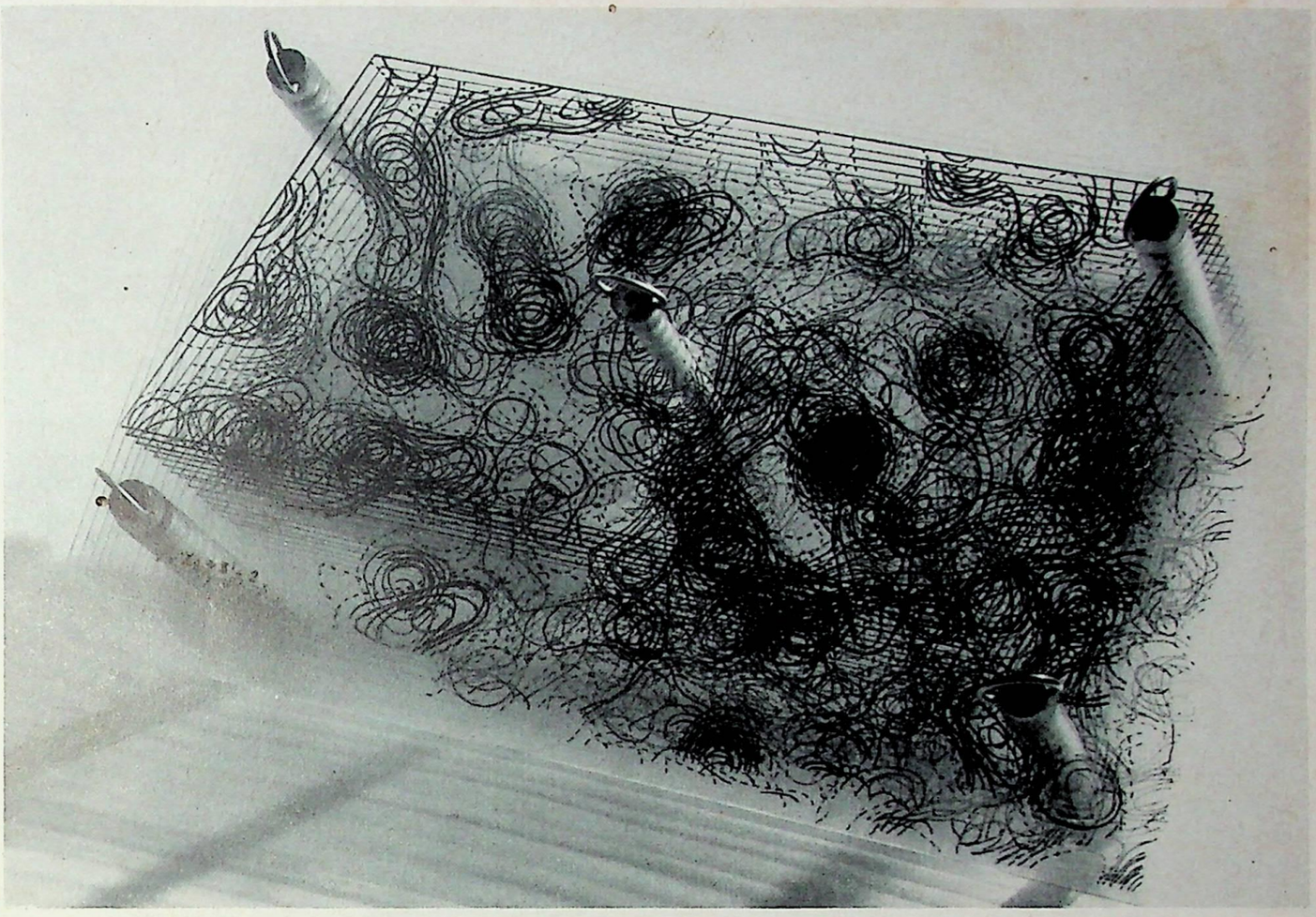


FIG. 5. Three-dimensional Fourier synthesis of the myoglobin unit cell. Some of the rod-like polypeptide chains can be seen.

present time, however, these theoretical limits of resolution are of academic interest only, and in fact, because the amount of labor goes up so rapidly with increasing resolution, we decided in the first instance to try out the isomorphous-replacement method at the lowest resolution that was likely to give us useful and interesting structural information.

How were we to decide what resolution would be useful? Here we had to bring in preconceived notions about protein structure, and we adopted a working hypothesis which would nowadays be accepted by most workers as being at least plausible—that α-helices are somehow involved in the protein molecule. There is some experimental evidence that this is the case, though it is not of an absolutely conclusive kind. An α-helix is a very dense object, and at low resolution it should appear as a rod of electron density about 1 electron/A^3. The side chains sticking out from the helix are more open structures, and they should appear as regions of density about 0.29 electron/A^3, surrounding the central rod. α-helices would be expected to pack together so that each is about 9 to 10 A from its neighbors. It was decided, therefore, to calculate the first Fourier synthesis at 6 A resolution, when the α-helices, if present, should be clearly distinguishable.

The process by which we finally arrived at the Fourier synthesis is not detailed here. Much of the calculation was done with a high-speed computer. The method is analogous to that used in two dimensions; it requires no guesswork, and the phases of all reflections can be cross-checked. We are, therefore, confident that our three-dimensional electron-density distribution is essentially correct, although it is somewhat blurred by its low resolution, and must contain a small background because of experimental error.

When the Fourier synthesis was finished, we had the problem of making a suitable geometrical representation of this three-dimensional object. We first plotted a series of electron-density contour maps for 16 parallel planes at different levels in the unit cell, and then traced the contours of these onto transparent sheets of Lucite, so that by stacking two or more of them together one could see how the contour surfaces were arranged in three dimensions (Fig. 5). At this point, we could see the two differently oriented myoglobin molecules which comprise the unit cell of the crystal, their positions being related to one another by the symmetry of the crystal (a screw dyad axis). We could also see rod-like features of high density within the molecules, and indeed except for a few regions of uncertainty we could follow a continuous path of high electron density throughout the entire molecule. Thus, we were able to make a model

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

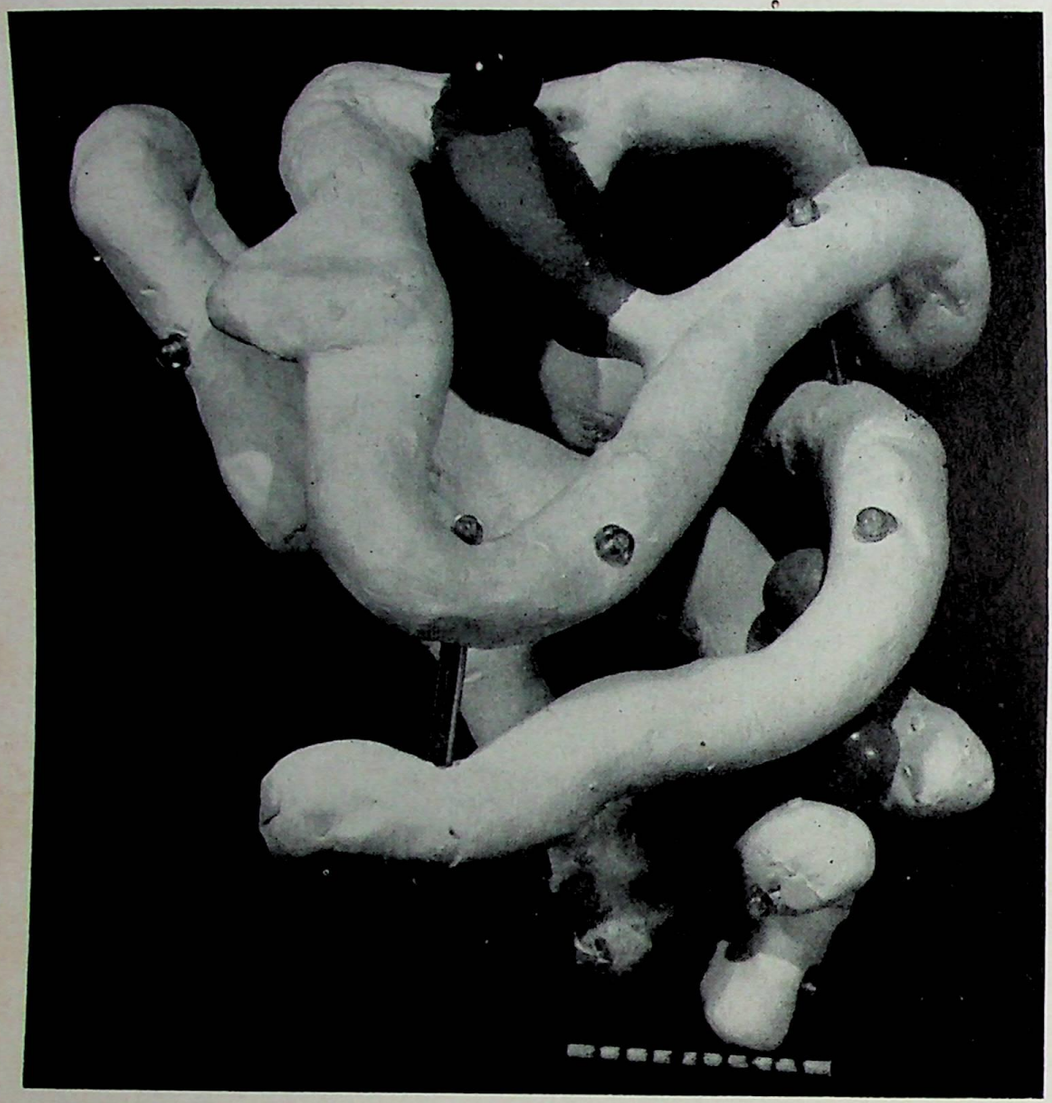


FIG. 6. Model of the myoglobin molecule. The grey disk seen nearly edge-on is the heme group; the small spheres are heavy atoms attached to the molecule. The marks on the scale are 1 A apart.

to represent the protein molecule consisting of one long rod of modeling clay or thermo-setting plastic, molded to the shape of the rod of high electron density. The rod is irregularly coiled to form a bundle measuring about 45 A by 35 A by 25 A. On looking at the model, one is impressed by its total lack of symmetry, and by the absence of parallel lengths of rod (Fig. 6).

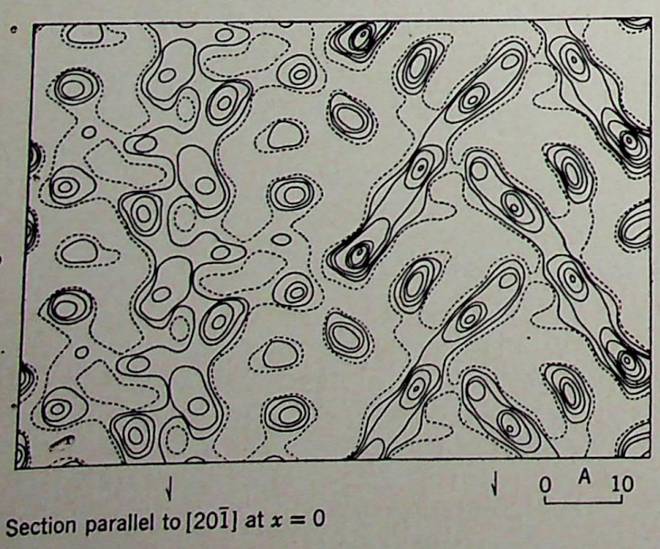


FIG. 7. Section through the three-dimensional Fourier synthesis showing, on the right, straight rods 40 A long.

The rod has segments over which it is nearly straight for distances of 20, 30, or even 40 A (Fig. 7). Adjacent segments of rods are always at least 9 to 10 A apart. The electron density at the rod axis is about 1 electron/A^3. Thus, in all respects, the rod seems to fit the specifications of an α-helix, and we have little doubt that it is in fact the polypeptide chain. We cannot yet be sure, however, whether the configuration of the chain is that of an α-helix, though it must be something fairly closely resembling this in gross dimensions.

From chemical evidence, the molecule is thought to consist of just one chain, and that is the way we represent it in our solid models. But, in following the high-density streak through the molecule, one finds regions where its path cannot be traced unambiguously. These are the places where we think the rod is sharply bent. The presence of a bend would imply an interruption in the very dense helical structure, and would explain our difficulty in tracing it through such a region. A simple calculation shows that, in any case, the entire chain

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

cannot be coiled as tightly into an α-helix, for the total length of rod appearing in our model is 300 A, whereas if the molecule consisted entirely of α-helix, the rod would be about 230 A long. If we make the obviously oversimplified assumption that part of the molecule has the α-helix form, and that the rest of it is fully extended, then the measured length of the rod requires that about 70% of it be in the helical form. This figure agrees with other estimates made by the optical-rotation method and by the deuterium-exchange method.

The heme group of myoglobin can also be seen in our Fourier synthesis. This group was easy to identify because of its iron atom, by far the heaviest atom in the protein molecule. It produced a peak of electron density 50% higher than any other feature in the molecule. Though we cannot resolve the porphyrin-ring system, the heme group appears flattened in one direction, and this direction corresponds very well with what we already knew of the orientation of the heme group from optical and magnetic studies of myoglobin crystals. In order to make our identification of this group more certain, we treated the protein with a reagent, *p*-iodophenyl hydroxylamine, which is known to react with the myoglobin heme group, and it turned out that the electron density was raised at the appropriate place in the molecule (Fig. 8).

We are now collecting further data in order to extend the resolution of our model to 2 A. At this resolution, our rod, if it is truly a helix, should show up as a hollow tube with spiral walls like a spring. We might even be able to recognize some of the larger side chains, such as those containing benzene rings. The resolution of individual atoms is still a long way off, if indeed it is attainable.

When considering the future of protein crystallography, one must remember that a full three-dimensional analysis of a protein crystal is a very long and tedious process. In that respect, it is like the determination of the amino-acid sequence of a protein. Neither is likely to be embarked on lightly, yet it is only by means of these now difficult and tedious methods that we will ever be able to put into concrete form the vague concepts concerning configurational specificity in proteins. Edmundson at the Rockefeller Institute has recently begun an amino-acid sequence analysis of myoglobin, and his work taken together with the x-ray results at higher resolution may eventually lead to a detailed picture of the configuration of this one protein. There are, however, very many other interesting proteins for which the same sort of dual approach would be highly warranted, and it is very important to try to develop short-cut methods, but as yet nobody has much idea what form they will take.

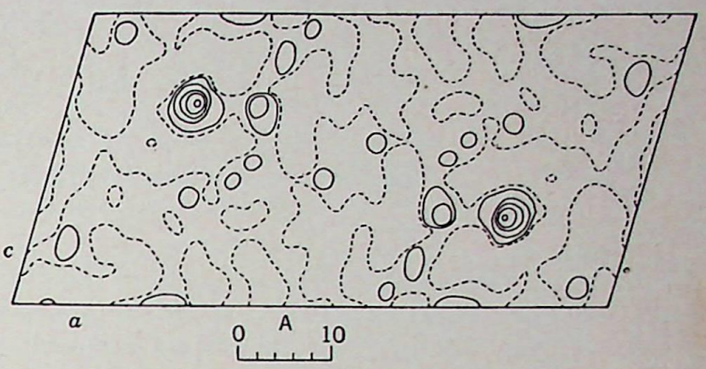


FIG. 8. Difference Fourier projection of the *p*-iodophenyl hydroxylamine derivative of myoglobin. The peaks are caused by the iodine atoms, and they lie near the heme groups in projection.

The author desires to express his appreciation of the work done by Dr. J. Kraut and Dr. R. G. Hart in preparing this manuscript for publication.

BIBLIOGRAPHY

F. H. C. Crick and J. C. Kendrew, Advances in Protein Chem. **12**, 133 (1957).

M. M. Bluhm, G. Bodo, H. M. Dintzis, and J. C. Kendrew, Proc. Roy. Soc. (London) **A246**, 369 (1958).

J. C. Kendrew, G. Bodo, H. M. Dintzis, R. G. Parrish, H. Wyckoff, and D. C. Phillips, Nature **181**, 662 (1958).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

10
The Hydrogen Bond

LESLIE E. ORGEL
Department of Theoretical Chemistry, University Chemical Laboratory, The University of Cambridge, Cambridge, England

IT is an experimental fact that a hydrogen atom attached to an electronegative atom in one molecule can cause an interaction with another molecule which also contains an electronegative group, or with a different electronegative group of the same molecule. This interaction is referred to as hydrogen bonding.[1] The order of hydrogen-bond strength is F>O>N≫Cl,C. There is no clear-cut lower limit to the strength of hydrogen bonds, those to chlorine and carbon atoms in particular giving rise to very little energy of interaction.

THERMOCHEMISTRY

The heat of formation of the strongest hydrogen bonds is about 10 kcal, but this value is achieved only in polymers of HF. The strongest hydrogen bonds between groups of biological interest are between oxygen atoms, and the heat of formation of these is rarely greatly in excess of 5 kcal. Hydrogen bonds between nitrogen and oxygen are usually somewhat weaker and bonds between pairs of nitrogen atoms somewhat weaker again. These relations apply only to general trends; it is not suggested that every O—O hydrogen bond is stronger than every N—O or N—N bond.

STEREOCHEMISTRY[2]

The lengths of hydrogen bonds vary considerably, but there is a general rule that the stronger the bond, the shorter its length. Oxygen bonds may have a length as short as 2.5 A, although those occurring in biochemical systems are usually somewhat longer than this, perhaps 2.7 A. There is no sharp upper limit to the length of a hydrogen bond, but if the O—O distance is much greater than 3.0 A, there is very little interaction. Similar considerations apply to N—O and N—N hydrogen bonds.

In almost all hydrogen-bonded systems, the proton does not lie symmetrically between the atoms bonded. The exceptions to this rule are the $[HF_2]^-$ ion and the hydrogen-maleate ion, but in systems of biochemical interest the hydrogen will always be associated definitely with one or the other of the atoms bonded.

The position of the hydrogen atom relative to the atom to which it is most strongly bound is always close to that which it would occupy if no hydrogen bond were formed. On the other hand, its position relative to the second atom involved in the bond is much more variable. Thus, both of the arrangements shown in Fig. 1 are encountered in certain proteins and in peptides. It is interesting that x-ray studies show that, when the general stereochemistry of a molecule makes it impossible for the hydrogen atom of a hydrogen bond both to lie on the line joining the atoms bonded and to occupy its normal position relative to its nearest neighbor, it is the second requirement which is satisfied. Thus, in salicylic acid the hydrogen atom lies well off the line joining the two oxygen atoms.[3]

SPECTROSCOPIC PROPERTIES

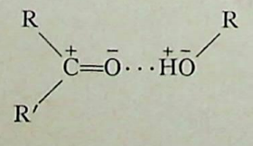
Hydrogen bonding has a profound influence, both on the infrared and the nuclear-resonance spectrum of a hydrogen-bonded system. Here it is mentioned only that these techniques are of great importance in studies of hydrogen bonding.

EFFECT OF HYDROGEN BONDING ON CHEMICAL PROPERTIES

In general, the effect of hydrogen bonding on the donor group is in the same direction as the effect of ionization, but less extreme. In a similar way, the acceptor group behaves in the same manner as it does when a proton is added, but again the effect is less marked. An example of this is salicylic acid which is a much stronger acid than one would have anticipated if no hydrogen bond were formed. The effect of the hydroxyl proton on the carboxylic-acid group is like that which would be produced by the addition of an extra proton and, therefore, leads to an easy loss of the carboxylic-acid proton.

THEORIES ABOUT THE HYDROGEN BOND

The earliest theories of hydrogen bonding postulated either an electrostatic or a covalent interaction, usually the former.[4] In the electrostatic theories, it is argued that, since the fluorine oxygen and nitrogen atoms are very electronegative, a proton attached to them must carry a positive charge. Thus, if a second molecule containing an electronegative and consequently negatively charged atom is available, the hydrogen atom will be attracted to it as shown below.

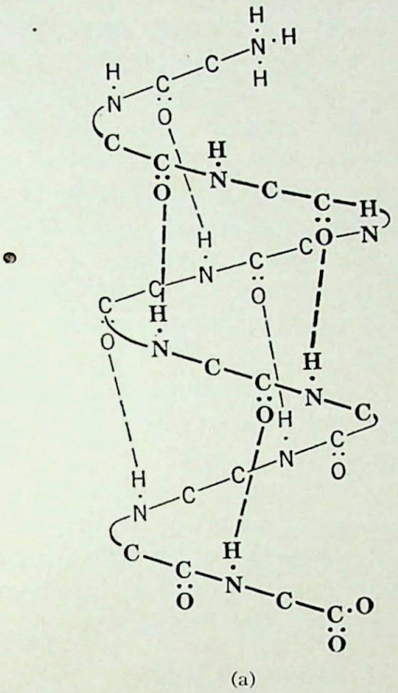


Naturally, the strength of the bond will be greater if the positive charge on the hydrogen atom is large and if the negative charge on the acceptor is also large. This explains why the strongest bonds are formed by fluorine, the most electronegative element, and the next strongest

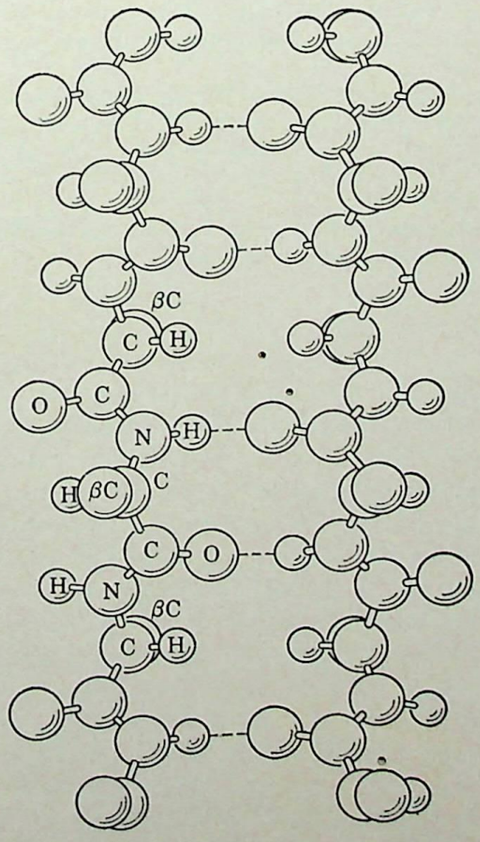
CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

by oxygen, etc. Rough calculations show that the energies of formation of hydrogen bonds are not incompatible with a simple electrostatic theory, while the covalent theory could not be tested quantitatively.

More-recent detailed calculations[4] show that the situation is intermediate between those suggested by the electrostatic and covalent theories. It seems that weak hydrogen bonds are always entirely electrostatic in origin, but that, as the hydrogen bond becomes stronger and shorter, the covalent contribution to bonding increases. In typical hydrogen bonds occurring in biological systems, it seems unlikely that the covalent character of the bonds will ever exceed about 10%.



(a)

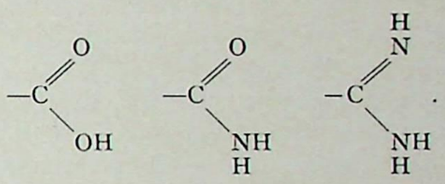


(b)

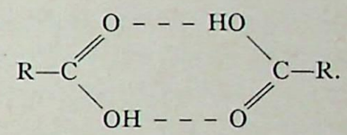
FIG. 1. Hydrogen-bond arrangement in (a) α-helix and (b) sheet structure.

PARTICULAR HYDROGEN-BONDING CONFIGURATIONS OF BIOLOGICAL INTEREST

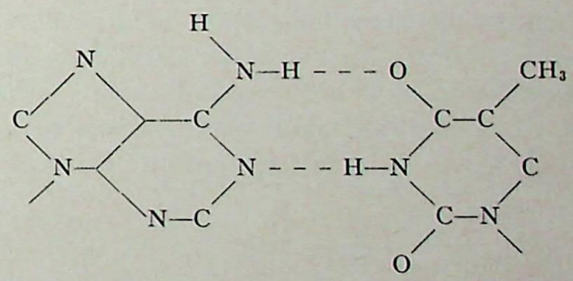
There is one arrangement of atoms which leads to particular stable associations between pairs of molecules. It is the one present in carboxylic acids, amides, guanidine, etc.



This arrangement is particularly advantageous since it allows for the formation of two bonds rather than one between molecules, e.g.,



We believe that it must be one of the most important contributors to the hydrogen-bond stabilization by side chains in proteins, etc., and it is also the basis of the Watson-Crick[5] base-pairing scheme, e.g., in the base pair adenine-thymine.



ELECTROSTATIC AND HYDROGEN-BOND INTERACTIONS

In proteins, one important interaction is that between charged amino groups and charged carboxylate ions. If the two ions are sufficiently close to be in contact, it is not possible to distinguish between electrostatic and hydrogen-bond contributions to the energy of interaction. Thus, it is only in the case of interaction between widely separated charged groups that a clear case of electrostatic (without hydrogen bond) interaction can be recognized.

HYDROGEN BONDING IN AQUEOUS SOLUTION

The heats of formation of hydrogen bonds in the gas phase give one little information about their stability in aqueous solution, because in this case one is concerned

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

with the difference in the hydrogen-bonding energy of the solute and solvent, firstly when the solute molecules are bound together and secondly when they are separated. Thus, the heat obtained on forming hydrogen bonds in solution is usually less than that obtained in the gas phase. Most simple single hydrogen bonds are split almost completely in water except when the solute is present at very high concentration. The amide group, however, because of its great hydrogen-bonding power, can form moderately stable dimers in aqueous solution.

Schellman[6] has made a detailed study of the dimerization of urea in aqueous solution. Although his assumptions are open to a good deal of doubt, the final value for ΔH, the heat of formation of an O—N hydrogen bond in aqueous solution of -1500 cal is probably roughly right. In the author's view, this figure is probably a little too high. Schellman then went on to calculate the degree of stability of the α-helix in aqueous solution. He was obliged to make a number of questionable assumptions and so, instead of obtaining a firm value for the free energy of formation, he could give only a rather wide range of possible values. His work is of great interest because it shows that in aqueous solution the α-helix must be on the border of stability so that side-chain interactions may be critical in determining whether a particular protein exists in the α-helix form or not. Even though the numerical values which he obtains are probably not very reliable, this conclusion—which was reached before much of the evidence which is described by Doty (p. 61) was obtained—is of very great importance.

Schellman discussed end effects in some detail and showed that the α-helix should be stable only provided the chain size exceeds a certain lower limit. Using reasonable values for the parameters involved, the critical length came out to between 8 and 15 units. Experimentally, there does seem to be some evidence for this conclusion.

PROTEINS IN AQUEOUS SOLUTION

The available experimental evidence discussed in detail in a later paper [Doty (p. 107)] shows that most proteins are only partially present in the α-helix configurations, the rest presumably being in some less regular structure. This raises a point of great importance to all discussions of the physical properties of proteins in aqueous solution, namely, the question of whether or not proteins are present in equilibrium configurations. It seems at least possible that, in fact, proteins are present in frozen configurations formed during the peeling-off of the molecule from its template. If so, many of the arguments from the structure of simple synthetic polypeptides may not apply directly to proteins. If this is true, any structural information about proteins also should give useful information about their mode of formation. In particular, it may be that certain sections of proteins which can be removed without effecting the enzymatic activity are present only in order to allow the protein to fold up in the right way. Of course, there are also many other possible explanations of the same effects. Experiments with synthetic copolymers containing, for example, appreciable quantities of glutamine or asparagine might help to solve this problem.

BIBLIOGRAPHY

[1] See L. C. Pauling, *The Nature of the Chemical Bond* (Cornell University Press, Ithaca, New York, 1940), for much introductory material.

[2] J. Donohue, J. Phys. Chem. **56**, 502 (1952).

[3] W. Cochran, Acta Cryst. **6**, 260 (1953).

[4] C. A. Coulson and U. Danielsson, Arkiv Fysik **8**, 245 (1954).

[5] F. H. C. Crick and J. D. Watson, Proc. Roy. Soc. (London), **A223**, 80 (1954).

[6] J. A. Schellman, Compt. rend trav. lab. Carlsberg. Sér. chim. **29**, 223, 230 (1955).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

11
Forces between Macromolecules

WALTER H. STOCKMAYER

Department of Chemistry, Massachusetts Institute of Technology, Cambridge 39, Massachusetts

IN earlier articles, special attention has been paid to electrostatic forces (see Rice, p. 69), steric forces (see Waugh, p. 84), and hydrogen bonds (see Orgel, p. 100) in biological systems. Here, the attempt is made to round out the discussion of various types of intermolecular action[1] and to take up several special applications.

One recalls first the nature of the mutual potential energy of two simple and chemically inert molecules, such as two argon atoms. Instantaneously, this energy is the sum of the Coulombic interactions involving all of the electrons and nuclei. However, one cannot follow the electronic motion in detail; all that is required is simply the average force between the nuclei, which move much more slowly. In fact, little error is made[2] by assuming the nuclei to be stationary while quantum-mechanical averages over the electronic motions are computed. The results of such theoretical calculations, reinforced by experimental evidence from properties of gases and molecular crystals, lead to a potential energy given approximately by relations such as:

$$\phi(r)=Ar^{-n} \quad -Cr^{-6}, \tag{1a}$$

or

$$\phi(r)=B\exp(-\alpha r) \quad -Cr^{-6}, \tag{1b}$$

where r is the internuclear distance and the other symbols are constants.* The positive (repulsive) term in these expressions is the steric energy resulting from direct overlap of the two atomic electron clouds, and the negative term is the dipole-dipole dispersion or London energy. The familiar curve for the function $\phi(r)$ is shown in Fig. 1(a).

Two further points concerning these intermolecular potential energies should be mentioned.

(a) In a system of many molecules, the total potential energy is to a good approximation the sum of those for all possible pairs; thus, useful predictions of the behavior of bulk matter can be made in principle with the aid of appropriate statistical-mechanical and kinetic theories, from a knowledge of the potential energy between just two isolated molecules. Without this fortunate result, progress would be difficult indeed.

(b) The most important features of the $\phi(r)$ curve—namely, the equilibrium distance and the potential energy corresponding thereto (the minimum in ϕ)—are not given either by steric or by London forces alone, but result from a *balance* between them. This truism warns that, in general, one should not hope to single out one unique type of interaction (and the classification into "types" is really somewhat arbitrary) as solely responsible for a particular effect.†

One of the simplest applications of a knowledge of $\phi(r)$ is found in the theory of slightly imperfect gases[3] for which the equation of state can be written

$$PV/RT=1+B(T)/V+\cdots, \tag{2}$$

where P is the pressure and V the molar volume. The so-called *second virial coefficient*, $B(T)$, is given for monatomic gases (except for the lightest molecules at very low temperatures) by

$$B(T)=2\pi N_0\int_0^\infty[1-\exp(-\phi(r)/kT)]r^2dr, \tag{3}$$

where N_0 is Avogadro's number and k Boltzmann's constant. At temperatures low enough so that the depth of the potential well is at least comparable to kT, the quantity $B(T)$ can be thought of as the negative of an equilibrium constant describing binary clustering of the molecules.

On turning to the interactions of two molecules *in solution*, as one inevitably must in most biological considerations, it is extremely important to recognize that the space intervening between the molecules is no longer empty, but is filled with solvent and perhaps other solutes. Detailed treatments of the interactions are now much more difficult, but Eqs. (2) and (3) may be retained if P now is defined to be the *osmotic pressure*, V the volume of solution per mole of *solute*, and if $\phi(r)$ in Eq. (3) is replaced by a quantity $W(r,P,T)$, called the *potential of mean force*.[4] The physical meaning of the latter quantity is easily stated: It is the work that would be required to pull apart the two solute molecules from a separation r to an infinite separation when these are immersed not in a vacuum (as for the potential energy ϕ), but in a very large quantity of the solvent, the molecules of which are permitted to perform all motions consistent with equilibrium at the stated temperature and pressure. Thus, W depends (through the properties of the solvent) upon pressure and temperature, and its dependence upon the separation r reflects the structure of the solvent.

The simplest and most familiar example of such a potential of mean force is the Coulombic energy

$$W(r,P,T)=e_ie_j/rD(P,T) \tag{4}$$

between two point charges, e_i, e_j, separated by a distance r in a medium of dielectric constant D.

* For pairs of simple molecules, n is about 12, or α about 3×10^8 cm^{-1}.

† An apparent exception is given by the Coulombic interactions in very dilute ionic solutions, but even here a short-range steric energy is essential to insure stability.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

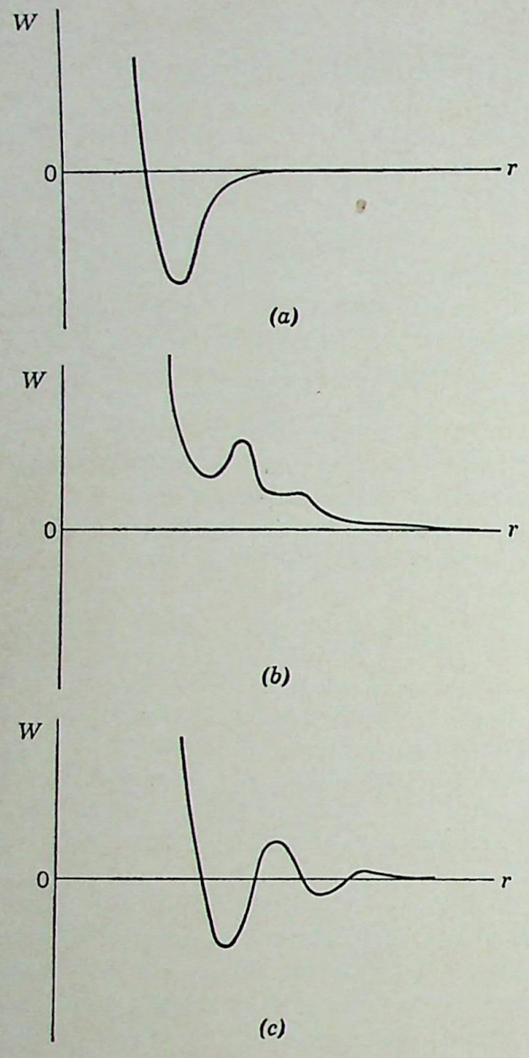


FIG. 1. Potentials of mean force $W(r)$ between two nonpolar spherical molecules as a function of their separation r: (*a*) *in vacuo;* (*b*) in good solvent; (*c*) in poor solvent.

Outside the domain of very dilute electrolyte solutions, such simple and general expressions for W as Eq. (4) are no longer found. Possible curves of $W(r)$ for two uncharged nonpolar solute molecules in a good and in a poor solvent, respectively, are drawn in Figs. 1 (b) and 1 (c). The oscillatory nature of these curves is imposed by the structure of the solvent, but it is seen that the damping is quite large. Thus, many approximate theories of liquids and solids can be constructed in which molecules are considered to interact only if they are nearest neighbors.[4]

The important role of the solvent has already been emphasized in connection with hydrogen bonding (see Orgel, p. 100). Perhaps unfortunately for students of both inorganic and biophysical chemistry, the structure of liquid water is probably uniquely complicated. As discussed in detail by Bernal and Fowler,[5] a large residuum of the characteristic, tetrahedrally coordinated structure of ice may be considered to persist in liquid water at ordinary temperatures. The orientations of the water molecules are far from random, and they oscillate torsionally with rather small amplitudes, instead of rotating freely. The introduction of solute molecules or ions into water produces large perturbations in the structure surrounding each solute particle, roughly out to about the third shell of neighbors. Despite considerable effort,[6,7] a complete quantitative theory of these effects still is lacking. However, the student of macromolecular behavior rarely needs an absolute theory, since his task is to infer the behavior of large molecules, given all necessary information about small ones.

Steric effects between macromolecules, in general, may be considered in terms of the well-known covalent and van der Waals radii of the constituent atoms.[8] Naive expectations may sometimes be misleading, as illustrated by the measurements by Chinai and coworkers[9] of the random-coil dimensions in dilute solution of a series of synthetic polymethacrylates

$$(-\mathrm{CH_2}-\underset{\displaystyle\mathrm{COOR}}{\overset{\displaystyle\mathrm{CH_3}}{\mathrm{C}}}-)_n$$

in suitably ideal ("theta") solvents.[10] The effective bond length, instead of increasing monotonically as R is varied from methyl to *n*-octyl, passes through a minimum at the *n*-butyl ester. Clearly, a detailed examination of the structure of the polymer-chain skeleton and its surrounding substituent groups and solvent molecules is necessary to the understanding of such behavior.

Turning now to a more detailed consideration of the London dispersion forces,[1,11] one first recalls that they may be considered as arising out of the interaction between instantaneously unsymmetrical electrical-charge distributions in the molecules concerned. Although a noble-gas atom such as argon has, on the average, a spherically symmetric charge cloud, an instantaneous snapshot of the atom, if such were possible, would reveal almost always an unsymmetrical array, with a dipole moment, and also, in general, higher moments of the charge distribution. If this unsymmetrically distributed set of charges is brought close to another atom, its field polarizes the latter into an unsymmetrical condition. The result of this mutual perturbation of the charge distribution is an attractive potential energy whose principal part comes from the interaction of the instantaneous dipoles. London's general formula obtained by second-order perturbation theory, is (for the case of two identical molecules)

$$\phi_{\mathrm{dis}}(r) = -\frac{3e^4\hbar^4}{4m^2r^6}\sum_k \frac{f_{0k}^2}{(E_k - E_0)^3}, \qquad (5)$$

where e and m are electronic charge and mass, h is the Planck constant divided by 2π, r is the internuclear distance, E_0 and E_k are the electronic energies of the isolated molecule in the general state and the kth excited state, respectively, and f_{0k} is the so-called *oscillator strength* of the $0-k$ transition, closely related to the intensity of absorption or emission of the corresponding radiation.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

It is seen that the dispersion energy is always attractive, and that it varies inversely as the sixth power of the intermolecular separation, in contrast to a dependence upon r^{-3} for two dipoles in fixed specified orientations. This indicates the second-order nature of the effect, as does the presence in Eq. (5) of properties related to the excited states of the molecules.

The quantities required for directly computing the sum in Eq. (5) are rarely available. However, useful estimates of dispersion forces can be made by appealing to a similar expression for the (low-frequency) molecular polarizability which contains a similar sum. If it is

$$\alpha = \frac{e^2\hbar^2}{m} \sum_k \frac{f_{0k}}{(E_k - E_0)^2}, \tag{6}$$

assumed (sometimes apparently without too great error) that the largest part of each sum is contributed by a single excited state, elimination between Eqs. (5) and (6) leads to the much simpler result

$$\phi_{\text{dis}}(r) = -\frac{3\alpha^2}{4r^6}(E_k - E_0), \tag{7}$$

in which the energy difference $(E_k - E_0)$ is of the order of magnitude of an ionization potential (say, 10 to 20 v).

The corresponding expression for the London dipole energy between two *unlike* molecules[1] shows that no great specificity is to be sought in these interactions. A rough rule, often embodied in approximate theories of solutions, is that the force between a pair of unlike molecules is the geometric mean of those between the related pairs of similar molecules, at the same value of r.

Equation (5) suggests that highly colored molecules, which have at least one large oscillator strength at a rather small excitation energy, should exert especially strong dispersion forces. This expectation seems to be supported[12] by the fact that many dyestuffs are extensively dimerized in solution at concentrations as low as 10^{-4} molar, often in spite of Coulombic repulsion between ionic charges (cf. Fig. 2). The interaction between lyophobic colloid particles (see Rice, p. 69, and Verwey and Overbeek[13]) is qualitatively similar, Coulombic repulsion at large separation giving way to London attraction at smaller separations until overlap repulsion produces the stable minimum.

Unless extensive arrays of conjugated double bonds or condensed aromatic rings are found in large molecules, their mutual London interactions can be regarded as the sum of contributions from all possible pairs of localized centers, which may be taken as the individual atoms or the bonds. This additivity is well substantiated in the properties of the normal paraffins, for example, and [recall Eq. (6)] also is found in refractivity. When nonlocalized electron orbitals are permitted by the molecular structure, the simple London formula, Eq. (5), is inadequate and unusual orientational dependence of the forces is found,[14,15] but the magnitudes are not unusually great (e.g., compare benzene with cyclohexane, or hexatriene with *n*-hexane).

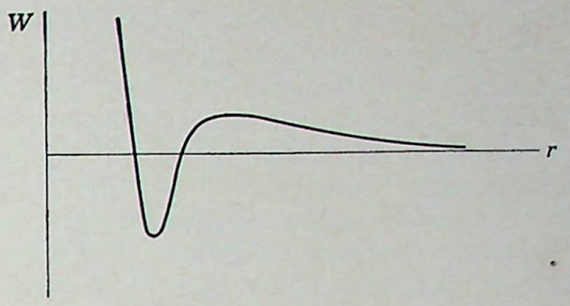


FIG. 2. Potential of mean force between two similarly charged molecules or lyophobic colloid particles. At large separations, Coulombic repulsion is dominant. As r decreases, this gives way in turn to London attraction and steric repulsion.

It has been suggested[16] that very strong and highly specific attractive forces can operate between identical macromolecules. Consider two identical and widely separated molecules, one in its ground state and one in some excited state. If these approach each other, the fundamental indistinguishability of the molecules prevents the identification of one of the two as the excited one. In familiar language, there is resonance[8] between two identical structures, with resultant (first-order) perturbation of the energy levels. The resonance energy (dipole term) varies as r^{-3} and can be either attractive or repulsive[1]; but, statistically, the former is favored at any finite temperature.

Resonance forces are highly specific indeed, for, if the excitation energies of the two molecules do not match rather closely, the resonance energy is negligible; but clearly, they cannot contribute much to the stability of a system unless appreciable numbers of the molecules are in excited states to begin with. This, in turn, requires low values of the excitation energies. The excited *electronic* states of all known macromolecules do not satisfy this requirement. Excited *vibrational* states of heavy molecules, on the other hand, are more generously populated but they do not make very large contributions to the forces (compare the magnitudes of the so-called "atomic" and "electronic" polarizabilities of molecules[17]), because only small changes of charge distribution attend the vibrational motions. It is concluded, therefore, that, although specific dipole resonance forces between identical molecules are certainly capable of existence, their magnitudes must be negligible in all practical cases. However, their role in energy-transfer processes involving excited states is very important.[18]

Another class of attractive force operates between electron acceptors and donors (Lewis acids and bases). These so-called "charge transfer" forces[19] contribute to the stability of many colored complexes (e.g., benzene-iodine) and, thus, have interesting consequences for spectroscopy, but they are probably too rare to be accorded major attention in biology.

These views regarding the forces between biological macromolecules are summarized by stating that, in general, they do not differ qualitatively from those between simple molecules, and that the most important specific interactions are steric and Coulombic.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

Turning now to applications, consider first the stability of micelles formed by aggregation of soap ions, which have long hydrocarbon tails and ionic heads. Whether the micelles are spherical [compare casein (see Waugh, p. 84)] or lamellar (as lipid membranes), the ions arrange themselves so that their charges remain at the micelle surface, where counterions from the surrounding medium congregate preferentially to complete an electrical double layer. The potential of mean force between unchanged hydrocarbon molecules in water is strongly attractive, so that the micelles would grow indefinitely large (as do droplets of uncharged hydrocarbon in water), except for the counteracting mutual repulsion between the ionic heads of the ions. More specifically, the standard free energy of formation ΔF° of a micelle containing n ions is the sum of a negative (favorable) bulk term directly proportional to n (arising from the net attraction between the hydrocarbon tails), and of a positive (unfavorable) surface term varying more rapidly‡ than n (arising from the electrostatic repulsion between the ionic heads). The resultant curve of ΔF° against n displays a minimum defining the most stable micelle size. Quantitative development of these ideas[20,21] accounts satisfactorily for the sizes and size distributions of soap micelles.

In systems of more than two components, electrostatic repulsion is not the only possible size-regulating mechanism. For example, the size of stable oil-in-water emulsion droplets is controlled by the available quantity of a third component, the surface-active agent, which need not be an electrolyte. Casein micelles (Waugh, p. 84) are characterized by large size and unusual stability toward added electrolytes, and may well involve the simultaneous action of several size-regulating effects.

Overbeek[22] suggests that lipid molecules naturally form flat membranes rather than spherical micelles because of the bulkiness of their hydrophobic parts (two hydrocarbon chains per glyceride unit), and that there is no need, therefore, to postulate a more specific biological mechanism of membrane synthesis. He also accounts for the order of magnitude of the observed spacing (50 to 100A) between multiple layers of such membranes (as in a myelin sheath) as a consequence of balance between London attraction and electrostatic repulsion.

The formation of gels or gelatinous precipitates involves a somewhat different balance of forces. Here, the interactions between the major portions of the primary particles, frequently chain- or rod-like macromolecules, are insufficient to cause aggregation and even may be weakly repulsive, but they are overshadowed by strong attractions between specific local sites. The resulting aggregate has a network structure and usually is extensively swollen by the solvent medium.

‡ The precise dependence of the electrostatic term upon n is complicated, and involves the charge and concentration of the counterions forming the double layer. If the counterion concentration is *very* low, this term varies as $n^{5/3}$ for spherical micelles or as $n^{3/2}$ for lamellar micelles.

The mathematical theory of gelation[23] is similar to that of explosions and of certain classical gambling problems. It may seem obvious perhaps that networks of unlimited size can form from primary molecules only if each carries an average of more than two reactive sites. The nature of the attraction between these sites need not be of any special type. Thus, in synthetic thermosetting resins,[23] the bonds are covalent and essentially permanent, while in gelatin[24] they are presumed to be the hydrogen bonds which stabilize the triple-stranded helical collagen structure. In antigen-antibody reactions,[25] a variety of possibilities exists.

If each of the primary molecules has a large number of reactive sites, only a very small attraction suffices to produce giant network structures. For example, if each primary molecule had 1000 reactive groups, a solution of about one percent concentration would gel even if the association equilibrium constant between two groups were only about 10^{-2} liter/mole. One can see that this mechanism affords the possibility of producing large effects (e.g., gelation greatly reduces the mobility of large particles) at small cost of free energy.

BIBLIOGRAPHY

[1] J. O. Hirschfelder, C. F. Curtiss, and R. B. Bird, *Molecular Theory of Gases and Liquids* (John Wiley and Sons, Inc., New York, 1954), particularly Chaps. 1, 12, 13, and 14.

[2] M. Born and J. R. Oppenheimer, Ann. Physik **84**, 457 (1927).

[3] See reference 1, Chap. 3.

[4] T. L. Hill, *Statistical Mechanics* (McGraw-Hill Book Company, Inc., New York, 1956).

[5] J. D. Bernal and R. H. Fowler, J. Chem. Phys. **1**, 515 (1933).

[6] R. W. Gurney, *Ionic Processes in Solution* (McGraw-Hill Book Company, Inc., New York, 1953).

[7] H. S. Frank and W.-Y. Wen, Discussions Faraday Soc. **24**, 133 (1958).

[8] L. Pauling, *Nature of the Chemical Bond* (Cornell University Press, Ithaca, New York, 1948), second edition.

[9] S. N. Chinai, J. Polymer Sci. **25**, 413 (1957).

[10] P. J. Flory, *Principles of Polymer Chemistry* (Cornell University Press, Ithaca, New York, 1953), Chaps. 12–14.

[11] H. Margenau, Revs. Modern Phys. **11**, 1 (1939).

[12] E. Rabinowitch and L. F. Epstein, J. Am. Chem. Soc. **63**, 69 (1941).

[13] E. J. W. Verwey and J. Th. G. Overbeek, *Theory of the Stability of Lyophobic Colloids* (Elsevier Publishing Company, Amsterdam, 1948).

[14] F. London, J. Phys. Chem. **46**, 305 (1942).

[15] C. A. Coulson and P. L. Davies, Trans. Faraday Soc. **48**, 777 (1952).

[16] H. Jehle, J. Chem. Phys. **18**, 1150 (1950).

[17] J. H. Van Vleck, *Electric and Magnetic Susceptibilities* (Oxford University Press, London, 1932).

[18] C. Reid, *Excited States in Chemistry and Biology* (Academic Press, Inc., New York, 1957).

[19] R. S. Mulliken, J. Am. Chem. Soc. **74**, 811 (1952).

[20] P. Debye, J. Phys. & Colloid Chem. **53**, 1 (1949).

[21] D. Stigter and J. Th. G. Overbeek, *Gas/Liquid and Liquid/Liquid Interface*, Proceedings of the Second International Congress of Surface Activity (Butterworths Scientific Publications, London, 1957), p. 311.

[22] J. Th. G. Overbeek (private communication, 1958).

[23] See reference 10, Chap. 9.

[24] P. Doty, Proc. Natl. Acad. Sci. U. S. **42**, 791 (1956).

[25] S. J. Singer, Proc. Natl. Acad. Sci. U. S. **41**, 1041 (1955).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

12

Configurations of Biologically Important Macromolecules in Solution

PAUL DOTY

Department of Chemistry, Harvard University, Cambridge 38, Massachusetts

ALMOST all biologically important macromolecules are particular chemical and structural modifications of three basic, polymeric-chain structures. These polymeric chains are the polysaccharides, the polypeptides, and the polynucleotides. They are better known in their more particular forms; for example, cellulose and amylose in the first instance, fibrous and globular proteins in the second, and deoxyribonucleic and ribonucleic acid in the third.

A polysaccharide chain such as cellulose exists as a completely extended chain in the crystalline, fibrous form and as such has a unique configuration. In solution, however, the limited rotational freedom permitted at the juncture of each repeating unit results in the chain being flexible and, because of internal Brownian motion, it undergoes continuous, worm-like changes in configuration. This state is referred to as a randomly coiled configuration. Such a configuration is devoid of any fixed relationship between pairs of residues, and hence is said to have no secondary structure. Since the fixed configuration in the crystalline form and the randomly coiled configuration in solution are typical of the situation found in most synthetic polymers as well, these have been widely studied and are well understood. Consequently, they are not further dealt with here.

The polypeptides and polynucleotides offer a sharp contrast in configurational properties to the polysaccharides and all other known polymeric chains. This uniqueness lies in the ability of these two types of chains to form hydrogen-bonded helical configurations consisting on one, two, or three chains which are stable in aqueous solution and hence in the cellular environment. Each of these helical configurations is unique and is equivalent to one-dimensional crystallites in that they consist of a periodic arrangement of the repeating-chain units along the helical axis. In this way unique, stereospecific relations in the individual macromolecule can be maintained while the macromolecule itself moves about in solution. The randomly coiled configuration exhibited in solution by all other macromolecules does not have this property. Consequently, it is evident that the preservation of unique configurations by polypeptides and polynucleotides in solution offers a basis of biological specificity.

These unique configurations of individual macromolecules are similar to the crystalline state in several respects, in addition to their having a one-dimensional periodic order. Most important is the implication that they will have a melting point; that is, a temperature will exist at which the supporting hydrogen-bonded structure will undergo a transition to the equivalent of the liquid state. For the macromolecule, the liquid state is simply the randomly coiled configuration already described. Thus, it is not surprising that transitions are found in these macromolecules that can have a sharpness approaching a phase transition. These are generally known as helix-coil transitions, and their study in the present case is of interest because their location reveals the relative stability of the unique configurations upon which one's attention becomes focused.

In the last few years, the study of these helical structures has greatly benefited from the possibility of making pure polypeptides and pure polynucleotides—pure in the sense that the repeating units are identical rather than differentiated as they are in the naturally occurring counterparts, proteins and nucleic acids. As a consequence, the properties of the purely helical forms could be carefully studied and the means of detecting such forms in the naturally occurring materials were thereby greatly sharpened.

It is against this background[1] that some studies of the configurations of polypeptides and their relation to protein structure are examined briefly, and following that a similar look is taken at the corresponding situation in polynucleotides and nucleic acids. The configurations have been established in solution by the use of the methods outlined earlier (Doty, p. 61). In addition, optical-rotatory dispersion and ultraviolet spectroscopy have been found widely useful in detecting the configuration established by the physical methods.

POLYPEPTIDES AND PROTEINS

In 1951, shortly after the proposal of the α-helix by Pauling and Corey,[2,3] Perutz[4] showed by means of x-ray diffraction that a few fibrous proteins in crystalline form did contain this configuration in unspecified amounts. However, it was not possible to extend this method to globular proteins.

In 1953, E. R. Blout and the author initiated a program of synthesis and characterization of synthetic polypeptides: this had as one of its aims the direct testing of whether or not polypeptides could take up unique configurations such as the α-helix in solution. In 1954, we found that poly-γ-benzyl-L-glutamate could exist in two configurations, the α-helix and the solvated, randomly coiled chain, depending on the solvent. Moreover, we found that the two forms showed a substantial difference in specific rotation similar in sign and magnitude

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

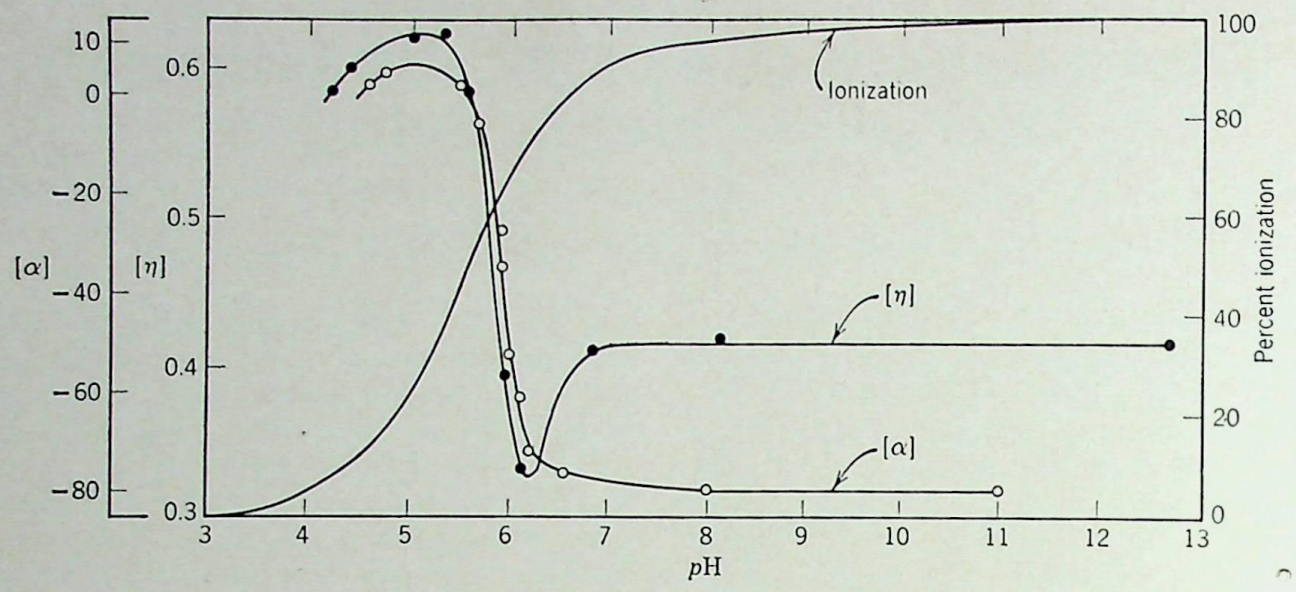


FIG. 1. The helix-coil transition in poly-L-glutamic acid.

to the difference between native and denatured proteins.[5] This fitted nicely with the suggestion made in 1955 by Cohen[6] that a specific main-chain configuration may be the cause of the difference in optical rotation between native and denatured states of proteins.

In 1956, more-detailed evidence was presented in support of the assignment of configuration in solution,[7] and it was shown that a sharp transition could be observed between the two configurations.[8] More to the point, however, was the investigation of the rotatory dispersion (i.e., the wavelength dependence of specific rotation), because this showed that, whereas the dispersion for the randomly coiled form was of the simple (or Drude) type found in most organic compounds containing asymmetric carbon, the dispersion of the α-helical form was anomalous.[8] At the same time, poly-L-glutamic acid was investigated in aqueous solution, and found to be helical in acid solution and randomly coiled in neutral and alkaline solutions.[9,10] This provided additional data on a system that more closely resembled proteins in aqueous solution.

It is perhaps useful to examine the case of poly-L-glutamic acid in some detail. Measurement of its intrinsic viscosity, sedimentation constant, and molecular weight[10] at *p*H values below *p*H 5 (where the carboxyl group was largely non-ionized) showed that it had the rod-like form and mass-to-length ratio of the α-helix. As the *p*H was raised, however, substantial changes in all of the physical properties of this polypeptide were observed[9] (Fig. 1), and the interpretation of these changes indicated that the polypeptide has gone over to a random-coil form. This can be considered as a melting process brought about by the electrostatic repulsion. That is, owing to the ionization of the carboxyl groups resulting in turn from the increase in *p*H, there is imposed upon the original helical structure an outer helix of negative charges; the repulsion between these charges is more than the hydrogen-bonding frame can withstand; and the α-helix structure breaks apart. One can observe this transition by measuring changes in the viscosity, for example, corresponding to the change in the shape of the molecule. However, the most generally useful of all of the changes accompanying this transition is the change in optical rotation. In this case, one observes a change of optical rotation from +4 to −80°C accompanying the helix-coil transition. With this nearly instantaneous means of detecting the configuration, it is easy to show that this is a reversible transition.

It is an old observation that the specific rotation of proteins falls* upon denaturation.[11] During the last decade, much specific support has been given to this generalization by quantitative studies, particularly by Kauzmann.[12] These studies show that the specific rotation, $[\alpha]_D$, for the majority of proteins lies between −30 and −60°C and that these values fall to approximately −100°C upon denaturation. Now, if the values of $[\alpha]_D$ observed for the helix and coil forms of poly-L-glutamic acid are adjusted to the mean residue weight of proteins (151 for sodium glutamate to 115 for proteins), the result is that a value of about +5°C would be expected for a protein in the completely helical form and about −105°C for a protein in the completely random-coil form. Thus, the observed specific rotation of proteins is compatible with 40 to 70% of their residues being in the α-helical configuration, and the specific rotation of denatured proteins corresponds to that of the randomly coiled configuration as previously surmised.[13] Naturally, this proposal requires extensive testing, and much of this is under way or has been done.

Since it is well known that proteins denature upon heating, one would expect the helical form to break up as the temperature is raised. The data in Fig. 2 show that this does occur if one uses the optical-rotation

* In this article, changes in specific rotation are always described with reference to the absolute value, not the magnitude as is customary. This is necessary since we now have both positive and negative values of specific rotation to discuss in contrast to earlier times when the only measured values of proteins were negative.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

method to gain a measure of helix content.[9] At *p*H 4.1 and temperatures below 40°C, the optical rotation is nearly constant. Above 40°C, the optical rotation begins to decrease, showing the start of the helix-coil transition. It melts out nearly all of the way at *p*H 4.6. One can pick up the rest of the transition by going to *p*H 5.0 where it melts out completely.

It is possible that the stability of the α-helix in poly-L-glutamic acid resides not so much in the hydrogen-bonding framework of the peptide units as in the pairwise interaction of carboxyl groups. This is the kind of hydrogen bonding that is always pointed out as being quite strong, and, if one makes models of the α-helix configuration of polyglutamic acid, one sees that uncharged glutamic residues can pair very nicely on the surface of the helix. Therefore, it is of great interest to see if other polypeptides that do not have this possibility also show the same phenomena.

Poly-L-lysine has an amino group on the side chain. This amino group is in the charged form ($-NH_3^+$) at *p*H values below 9.5, and becomes uncharged ($-NH_2$) above *p*H 10.5. Thus, one would expect that, if the helix is only marginally stabile, it would exist only at high *p*H values and would melt out as one lowered the *p*H. Current experiments show just this.[14] Again, there is about a 90°C change in optical rotation upon passing through the helix-coil transition. Physicochemical investigations were used to show that it exists as a pure α-helix at high *p*H.

If one examines the state of charge of the poly-L-lysine molecule when the helix begins to melt out in a quantitative fashion, one sees that this helix is considerably weaker than the poly-L-glutamic-acid helix. However, the difference is not enough to prevent the poly-L-lysine helix from being stable in aqueous solution. From this and a few other examples, it can be concluded that the α-helical configuration is stable in aqueous solution in the absence of pronounced electrostatic repulsions or other disrupting influences.

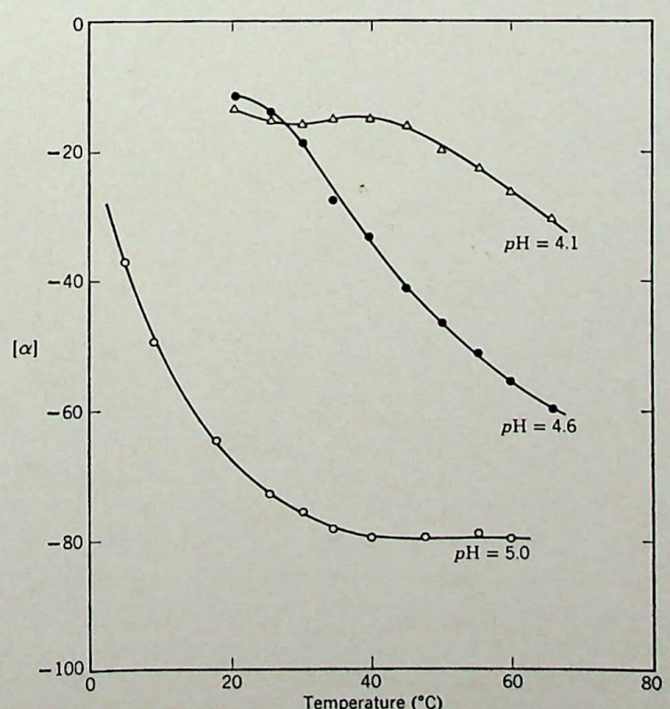


FIG. 2. Optical rotation of poly-L-glutamic acid as a function of temperature as several values of *p*H.

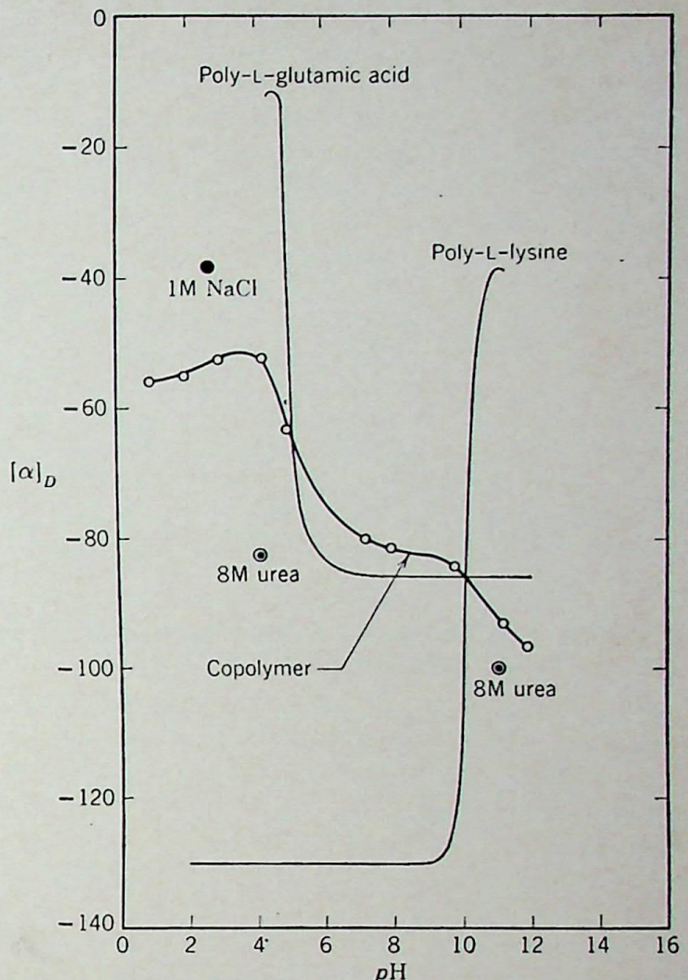


FIG. 3. The specific rotation as a function of *p*H for copoly-L-lysine-L-glutamic acid.

An interesting point associated with the nature of these helix-coil transitions is the question of whether or not they have an "all-or-none" character. That is, at the midpoint of the transition are half of the molecules in each configuration or are the residues in each molecule partitioned between these two forms? One of several answers indicating that the latter was the case was obtained in the following way. The intrinsic viscosity increases much more strongly with molecular weight in the helical form than in the random-coiled form. Thus, there is one molecular weight for which the viscosity is the same in both configurations. By choosing a sample of this molecular weight and carrying it through the transition, the intrinsic viscosity would be expected to remain unchanged if the transition were of the all-or-none type. This was not observed in several cases.[10,14,15] Consequently, the transition must be viewed as one that proceeds via intermediate states in which individual molecules have interspersed helical and nonhelical regions.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

It is of some interest to consider the behavior of a copolymer composed of equal amounts of L-glutamic acid and L-lysine. The optical rotation, which again is taken as a measure of helix constant for this copolymer, is shown in Fig. 3.[16] These results can be compared with those previously reported for poly-L-glutamic acid and poly-L-lysine which are also shown. The copolymer exhibits an intermediate behavior here; that is, it is about 50% helical at acid *p*H, and not at all helical at alkaline *p*H. This is interpreted as resulting from the difference in intrinsic stability of the glutamic-acid and lysine residues in supporting the helical structure. When the lysine residues are uncharged, there appears to be no helix, because the lysine residues cannot compensate for the electrical repulsion arising from the 50% negative charges that exist here owing to the ionized glutamic-acid residues. At neutral *p*H, one finds about 15 or 20% helix.

ROTATORY DISPERSION AND HELICAL CONTENT OF POLYPEPTIDES AND PROTEINS

Although one could proceed to estimate the fraction of residues (helical content) in the helical configuration from the specific rotation of proteins in the manner indicated in the foregoing, one would be failing to take advantage of another aspect of the situation that, upon proper analysis, offers an independent means of estimating the helical content and in addition offers a much more perceptive view of the problem. Soon after the first observations of the helix-coil transition in polypeptides, the rotatory dispersion (wavelength dependence of the specific rotation) of the two forms was measured[8] and found to be simple Drude dispersion for the random-coil form and a complex dispersion for the helical form. The complex dispersion could be fitted in a manner expected for coupled oscillators and originally suggested by Moffitt.[17] This gave rise to a very considerable theoretical interest in the problem of the rotatory dispersion of helical macromolecules,[18–22] but it has not been possible to take advantage of much of this because of the limited number of parameters that can be uniquely determined from the experimental data.

From the practical point of view, it is sufficient to say that the following equation was found adequate to express the observed data[23–25]:

$$[m'] = \frac{3}{n^2+2}\frac{M_0}{100}[\alpha] = (a_0{}^R + a_0{}^H)\left(\frac{\lambda_0{}^2}{\lambda^2-\lambda_0{}^2}\right) + b_0\left(\frac{\lambda_0{}^2}{\lambda^2-\lambda_0{}^2}\right). \quad (1)$$

Here, the specific rotation is multiplied by two factors that eliminate the trivial effect of residue weight, M_0, and refractive index, n. The corrected specific rotation, called the effective residue rotation, $[m']$, is then assumed to be made up of two components. The first of these is a Drude term but with two parts to its coefficient, one depending on the intrinsic residue rotation (asymmetric carbon), $a_0{}^R$, and the second is owing to the effect of the helical configuration, $a_0{}^H$. The second term is the anomalous one, being the square of a Drude term, and its coefficient is wholly owing to the effects of the helical configuration on the rotatory dispersion.

This equation is of the same form as that proposed by Moffitt and shown to be applicable to the first dispersion data obtained on helical polypeptides,[8,9,20] but it now seems proper to consider it an empirical proposal for two reasons. The equation itself is of the form expected for coupled oscillators. Secondly, Moffitt concentrated his attention on the second term, recognizing the complexities that may be associated with the first since, in any rigorous view, the first term must instead be replaced by several Drude terms each having a different λ_0. But if this last point is admitted, then the analysis of any available dispersion data is impossible, because there is not enough precision to allow the independent assignment of values to more than the three constants contained in Eq. (1), i.e., $(a_0{}^R+a_0{}^H)$, b_0, and λ_0. Consequently, in order to proceed, it was necessary to assume that λ_0 had the same value in both terms. This compromise, added to the later finding that the absolute value of b_0 cannot be computed with enough accuracy to be useful, forces us to accept Eq. (1) as empirical. It is important to make this point because, when it is recognized that Eq. (1) is not firmly supported by theory, it is more likely that it will be used with the caution that an empirical relation deserves.

The Fitts-Kirkwood theory[18,19] cannot be expressed in simple functional form, but since it consists of four different contributions, it is clear that these cannot possibly be resolved from experimental dispersion data since again too many parameters would have to be independently evaluated. However, if one takes the computed results of Fitts and Kirkwood,[22] one finds that they can be quite well fitted by Eq. (1). This observation emphasizes that the procedure that we have adopted in using Eq. (1) is not in conflict with the Fitts-Kirkwood theory.

Examine the adequacy of Eq. (1) in representing the dispersion data of polypeptides in the helical form. By dividing Eq. (1) by $\lambda_0{}^2/(\lambda^2-\lambda_0{}^2)$, it is clear that dispersion data obeying this relation should yield linear plots when $[m'](\lambda^2-\lambda_0{}^2)/\lambda_0{}^2$ is plotted against $\lambda_0{}^2/(\lambda^2-\lambda_0{}^2)$, provided the appropriate value of λ_0 has been selected. In this plot, the intercept of the straight line would then provide the value of $(a_0{}^R+a_0{}^H)$, which may be denoted a_0 for convenience, and the slope would provide the value of b_0. Since in this type of plot an empirical relation is being tested, it will be necessary to have specific rotation measurements at a number of wavelengths, because the absence of curvature in the plot must be demonstrated in each application.

In selecting the appropriate value of λ_0, a process of trial and error must be followed until that particular

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

value which linearizes the dispersion data is found. In the fitting of the first dispersion data, a value of 212 mμ was found for poly-γ-benzyl-L-glutamate and poly-L-glutamic acid. Since that time, the value has been confirmed on both of these polypeptides[10,26] and found to apply to at least four other helical polypeptides, as well as to several copolypeptides.[27] Dispersion data plotted in this manner are shown in Fig. 4 for three samples of poly-γ-benzyl-L-glutamate.[26] One sample had an average chain length of 4 units and gave a horizontal plot indicative of no helical content. Another had an average chain length of 10 residues and gave an intermediate slope and intercept, which, in comparison with the high molecular-weight sample, indicates about 60% helical content.

In all of the cases referred to in the foregoing (poly-γ-benzyl-L-glutamate, poly-L-glutamic acid, poly-carbobenzoxy-L-lysine, poly-L-lysine, and the copoly-L-lysine-L-glutamic acid), the values of b_0 obtained from the slopes of what one might call the "coupled oscillator" plot have clustered about −630. Thus, there is good reason to accept this value with an uncertainty of no more than ±10% as a constant characterizing in an empirical way the anomalous dispersion of helical polypeptides of L-residues. In some cases, we have found somewhat lower values, but these have in each case been with polypeptides that had not been proved to be in the completely helical form.

Turning finally to the coefficient a_0, we do not expect to find constant values for different polypeptides, because a_0 consists of two components, one of which (a_0^R) depends on the intrinsic rotatory power of the individual amino-acid residue. It is sufficient to say that, in the cases of the five polypeptides mentioned in the previous paragraph, the values of a_0 gave an average of zero. Since in the helical form a_0 is found to equal 650, this must be approximately the value of a_0^H itself.

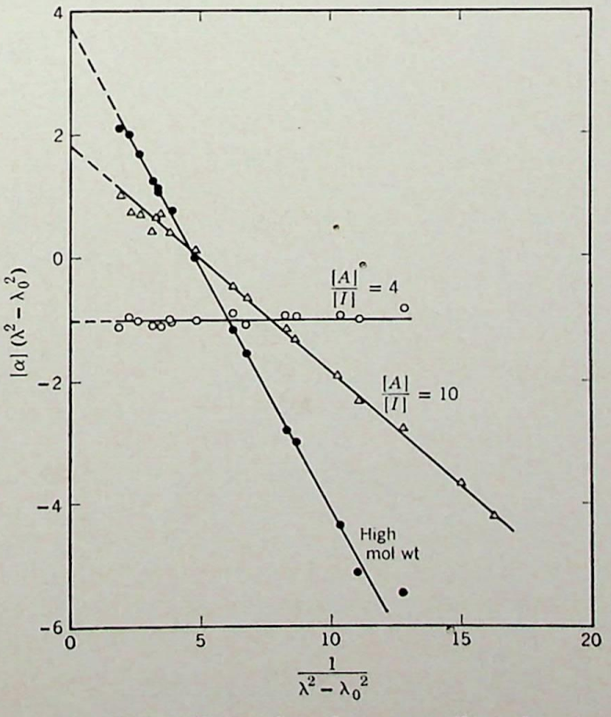


FIG. 4. Coupled oscillator plot of rotatory-dispersion data for poly-γ-benzyl-L-glutamate of various molecular weights. [A]/[I] indicates degree of polymerization (number average).

One is now in a better position to examine the extent to which the work on polypeptides enables one to understand the optical activity and rotatory dispersion of proteins. Actually, one is forced to proceed in a very simple manner. The polypeptide studies have given quantitative information about the rotatory dispersion of two configurations of polypeptide chains. One can, therefore, only ask if the rotatory dispersion of proteins can be accounted for by a linear combination of these two characteristic rotatory dispersions. It is important to recognize that, if this is the case, the residues not occurring in the α-helical configuration need not be in a randomly coiled configuration. It is necessary only that they not be in any kind of periodic arrangement, because this is the condition that the rotatory dispersion be of the kind observed for the randomly coiled configuration, subject of course to "solvent effects" on the intrinsic residue rotation, i.e., a_0^R.

The data summarized in terms of a_0^H and b_0 values can be translated directly into the contribution that the α-helical configuration would make to $[\alpha]_D$. It is +117° for the values of $a_0^H=650$, and $b_0=-630$ if the mean residue weight is taken as 115. The value of a_0^H contributes 84% to this result and, hence, it is the uncertainty of this term that affects most the predicted value. Taking the limits of a_0^H to be 550 and 750 would make the helix contribution to $[\alpha]_D$ 17° higher or lower, respectively. Thus, the prediction is made that for proteins the purely helical configuration and the completely denatured form should be separated by 100° to 135° in specific rotation. The polypeptide studies, adjusted to mean residue weights on 115 and aqueous solutions, establish the lower end of this increment—that is, the denatured protein—at about 110° and the high end—the pure helix—at about +5°. This is nearly the same as that obtained earlier from the measurements on poly-L-glutamic acid.

In terms of the model proposed here, the rotatory dispersion of proteins in aqueous solution should be given by converting Eq. (1) into the following form[28]:

$$[\alpha]=1.39\left[(\sum a_0^R+fa_0^H)\left(\frac{\lambda^2}{\lambda^2-\lambda_0^2}\right)+fb_0\left(\frac{\lambda^2}{\lambda^2-\lambda_0^2}\right)\right],$$

where f denotes the fraction of residues in helical form and a_0^R the sum over the a_0^R values characteristic of each residue in the protein. The model is not subject to a considerably more rigorous test than in the comparison dealing only with $[\alpha]_D$. This can be carried out in the following way. Rotatory-dispersion measurements are made on the protein both in aqueous solution and in the completely denatured state. For the former set of data, $[\alpha]1.39(\lambda-\lambda)$ is plotted against $(\lambda-\lambda)^{-1}$. The result should be a straight line yielding a slope equal to fb_0

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

TABLE I. Excess right-handed helical contents of (f) of various proteins in water.

	$-b_0/630$	$a_0^H/650$
Tropomyosin	0.88	0.87
Insulin	0.38	0.57
Bovine serum albumin	0.46	0.58
Ovalbumin	0.31	0.50
Lysozyme	0.29	0.39
Pepsin	0.31	0.26
Histone	0.20	0.30
Ribonuclease	0.16	0.17
Globin (H)	0.15	0.09

and an intercept equal to $(\sum a_0^R + f a_0^H)$. With the value of b_0 set at -630, the value of f, the fraction of residues in the helical configuration, is obtained at once.

The rotatory dispersion in the denaturated state should yield at once the value of a_0^R, since $f=0$ in this case. Indeed, only a measurement of the specific rotation at one wavelength is needed if it is known that $\lambda_0=212$ mμ, as appears to be the case in 8M urea. With a_0^R known, the value of fa_0^H can be obtained from the intercept evaluated in the plot described above. By taking the value of a_0^H as $+650$, an independent estimate of f can be made.

This kind of analysis has been applied to a number of proteins.[28] In each case, a linear plot was obtained. The test of the model consists then in seeing if the two different estimates of f are reasonable, that is, lie between 0 and 1, and are in agreement. A selection of this data is shown in Table I. It is seen that the values of f are indeed reasonable ones and that the agreement is fairly good.

The reasonably good performance of our model in this test is not, of course, proof that the model is a faithful description of the secondary structure of the proteins to which it has been applied. However, it does appear to represent an advance which deserves further and more-incisive testing. At present, it greatly increases our confidence in the existence of regions having the α-helical structure in proteins, it sharpens our views on protein denaturation, and it provides a framework of reference in which studies relating protein structure and function can be at least tentatively interpreted.

Having observed the striking dependence of the configuration of polypeptides on solvent, we were led to wonder if the intermediate development of the helical structure of proteins suggested by the foregoing experiments could not be increased by altering the solvent. For this, a solvent that was miscible with water, of comparable polarity and cohesive energy density, but with less hydrogen-bonding capacity, was needed. Our search indicated that 2-chloroethanol was well suited, and it was found that the addition of this to aqueous solutions increased the helical content as measured by rotatory dispersion in nearly every case. Some examples are shown in Table II for proteins dissolved in chloroethanol.[28]

Thus, the expectation that the helical content of proteins can be increased by lowering the hydrogen-bonding capacity of the solvent is confirmed. This behavior is precisely the opposite of denaturation. In the few cases that we have studied in detail (e.g., ribonuclease and insulin), the configurational changes on going from water to chloroethanol and back to water have been completely reversible. Thus, it would appear that the helical content of proteins is a result of balance between the intramolecular hydrogen bonding of the protein and the hydrogen-bonding capacity of the solvent. When water is the solvent, the point of balance is in many cases near 50%, and can be shifted in either direction by proper alteration of the solvent.

It is of little interest to note that insulin was the only protein in the list whose helical content was not substantially increased in chloroethanol over water. Since the model of insulin proposed by Lindley has about 20% of the residues in the left-handed helical configuration, owing to restrictions imposed by the cystine bridges, it appears quite possible that insulin has approximately this configuration both in aqueous and in chloroethanol solutions. The residues in the left-handed helix cancel the effect of an equal number in the right and thus, even when the helical content is nearly 100%, only about half of this amount registers with the rotatory-dispersion method. This anomaly is removed when it is recognized that the method measures only the excess right-handed helical content, as indicated in the titles of the tables. Nevertheless, this is the sort of detail that emerges only on making additional measurements. Obviously, there is now a great need of a completely independent method to measure total helical content so that the rotatory-dispersion methods can be checked in regularly behaving cases, and the amount of left-handed helical configuration evaluated in those proteins where cystine bridges may prevent the normal development of right-handed helical configurations.

As a final illustration of the relation between rotatory dispersion and protein structure, no better example can be quoted than the current work of Kay and Bailey on *Pinna* tropomyosin.[29] Kay's study of this molecule by physical methods shows it to have the character of thin, rigid rods having the dimensions of an α-helix of the proper molecular weight. Rotatory-dispersion measurements have now shown it to behave precisely like an α-helix with values of a_0^H and b_0 equal to those used

TABLE II. Excess right-handed helical contents of various proteins in chloroethanol.

Tropomyosin	110
Insulin	45
Bovine serum albumin	75
Ovalbumin	85
Lysozyme	63
Pepsin	44
Histone	72
Ribonuclease	67
Globin (H)	74

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

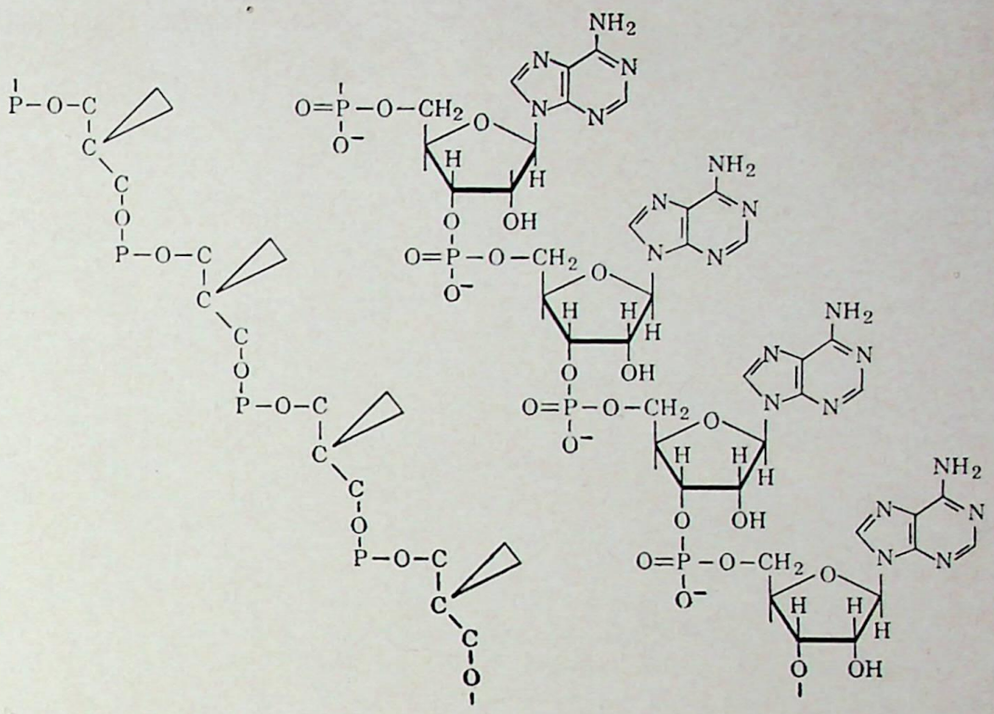


FIG. 5. The chemical structure of polyriboadenylic acid.

here.[30] Consequently, one has here the perfect example of the extreme of the scale of protein structure provided by our model. However, the work that is needed to establish this model as a framework of reference for protein structure in general remains to be done for the most part. It is certainly to be expected that numerous exceptions will be found, but there is even now reason to hope that the concept of describing the secondary structure of proteins by a partitioning of the residues between helical and nonhelical regions will play a useful role in the advance of knowledge of protein structure.

POLYNUCLEOTIDES AND NUCLEIC ACIDS

From the viewpoint of chemical structure, nucleic acids are known to consist of chains of repeating units such as are shown in Fig. 5. The backbone is a phosphoester polymer with six chain atoms per repeating unit. The chain shown in Fig. 5 is of the ribonucleic-acid type. That of deoxyribonucleic acid differs in that the oxygen at the 2-position on the ribose ring is absent. Generally, four different monomeric units, differing in the heterocyclic rings attached to the ribose group, are found in a given nucleic acid. In ribonucleic acid (RNA), these four groups are adenine, uracil, guanine, and cytosine. In deoxyribonucleic acid (DNA), thymine (5-methyl uracil) replaces uracil. DNA is found in cell nuclei as the principal component of chromosomes, and plays the central role in carrying and passing on the genetic endowment of the cell. RNA occurs principally at the sites of protein synthesis in the cytoplasm, and is intimately involved with that process.

Deoxyribonucleic Acid

The nucleic-acid chain is a very flexible one and, in the absence of secondary structure, it would have the configuration of a random coil. About 10 years ago, it became evident that this could certainly not be the case for DNA. With increasing refinement in the study of DNA in solution, it became clear that this molecule was the most extended ever examined.[31] Light-scattering studies and, more recently, viscosity and sedimentation measurements, show that typical DNA samples have average molecular weights in the range of 5 to 10 million and occupy volumes in solution about one-half micron in diameter. Ordinarily, polyelectrolytes are characterized by their molecular size being very dependent on ionic strength. This was not the case for DNA. Thus, it appeared to be not only very greatly extended, but stiff rather than flexible, as well.

These observations, as well as a number of others, became understandable with the structural proposal made by Watson and Crick[32] in 1953 on the basis of the x-ray diffraction studies made by Wilkins and his collaborators.[33] This structure consisted of two antiparallel DNA chains united through hydrogen bonds connecting the heterocyclic rings (usually called bases). Only two types of pairings were permitted, those between adenine and thymine and those between guanine and cytosine as shown in Fig. 6. The resulting structure is a two-stranded helix about 20 A in diameter. With two residues every 3.4 A, a DNA molecule with a molecular weight of 10 million would have a length of 50 000 A (5 μ). Such a long, thin structure would have a slight flexibility, presumably enough to account for several gentle folds that would reduce its maximum extent by about tenfold to agree with the molecular size found in solution. (Technically, it is correct to speak of this molecule as randomly coiled, but the degree of coiling is minute as compared with the single-chain coils to which this description is usually applied.)

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

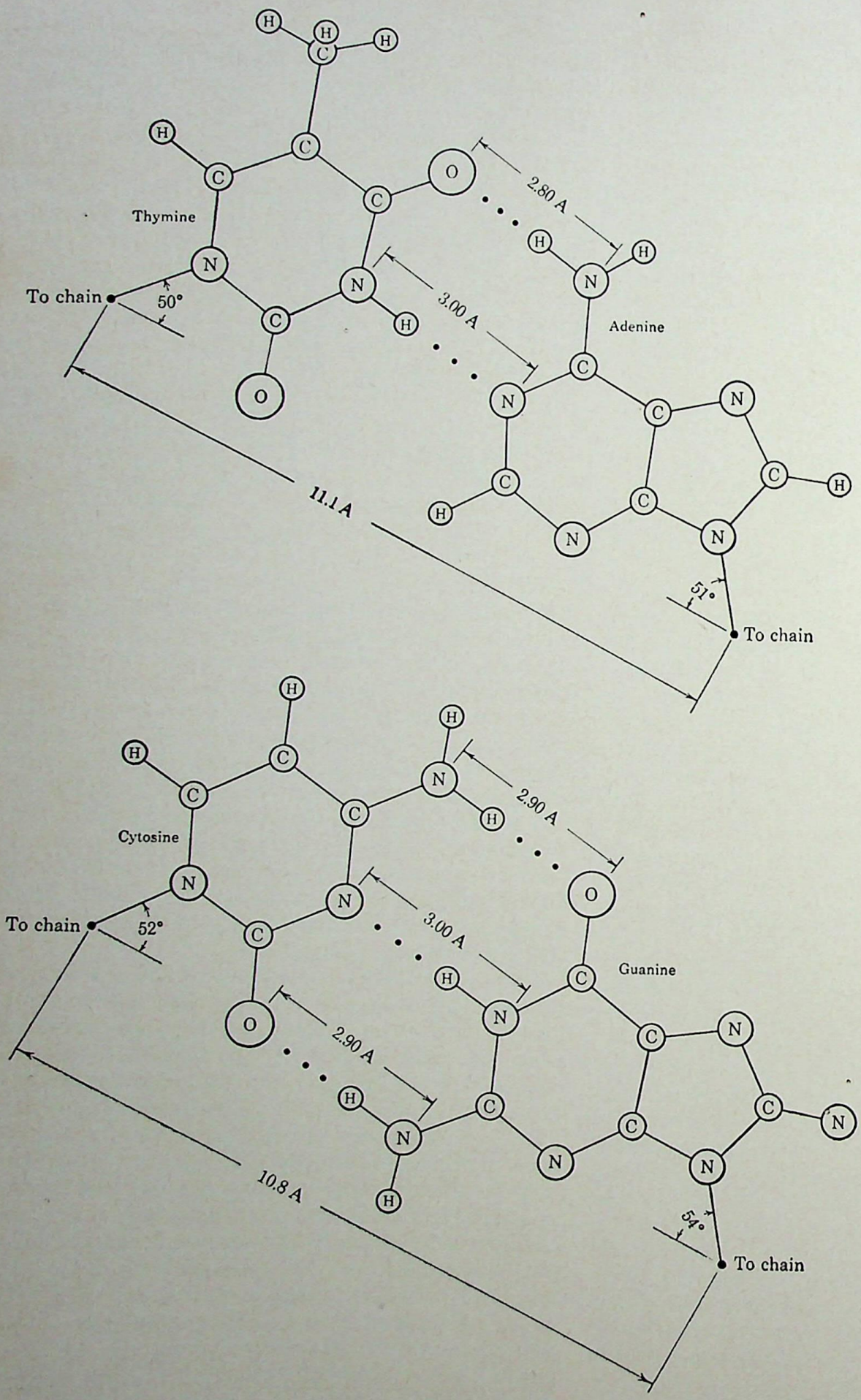


FIG. 6. Detail of the hydrogen-bonded base pairs occurring in DNA.

This periodic, secondary structure undergoes a cooperative melting out upon raising the temperature. The transition can be seen in terms of the viscosity, since the collapse of the structure is accompanied by a twenty-fold decrease in intrinsic viscosity. This is illustrated in Fig. 7.[34] The transition in this case is essentially irreversible. The reason for this is the low probability of the bases to find their original partners: without this the

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

perfect arrangement in the original DNA cannot be recovered. This helix-coil transition can be induced by other means, for example by raising or lowering the *p*H, and it can be followed by other means, such as changes in optical rotation. However, the most sensitive indicator of structure has turned out to be the ultraviolet spectrum. DNA exhibits a broad maximum at about 2600 A, and it has been widely observed that the maximum of this absorption is substantially depressed relative to that found for the corresponding monomeric units (nucleotides). Thus, hydrolysis which brings about such a conversion from DNA in the helical form to individual nucleotides is accompanied by approximately a 40% increase in the optical density or extinction at 2600 A. This suppression of absorbance in the DNA is known as hypochromicity. Now, it is found also that, if the helical form is converted to the randomly coiled, denatured form by raising the temperature or by lowering the *p*H, a rise of nearly 40% in the extinction coefficient also occurs. Thus, electronic states of the base groups are substantially different, depending on whether or not they are held in a hydrogen-bonded, helically arranged configuration. It is not yet clear if the hypochromicity is a result of the hydrogen bonding or of the stacking of the base groups one on top of another.

Polynucleotides

In 1955, Ochoa and Manago discovered a new enzyme which could bring about the polymerization of nucleoside diphosphates to form polymers having the chemical structure of ribonucleic acid. The study of pure polynucleotides which this has made possible has greatly increased knowledge of the basic properties which are combined in naturally occurring RNA itself. The problem that can be directly attacked with these new polyribonucleotides is the following. In 1956, Donohue[35] showed that the two specific base pairs used in the Watson-Crick structure of DNA are only two out of a large number of possible pairs. Thus, the availability of these polymers has made possible the investigation of which pairs are actually stable in aqueous solution. Warner[36] and Rich[37] quickly showed that polyadenylic acid (Poly A) and polyurdylic acid (Poly U) combined to form a double-stranded helix when their solutions were mixed. A number of other pairs have now been found, and in addition several triple-stranded helical complexes have been demonstrated. These are discussed elsewhere by Rich (p. 191).

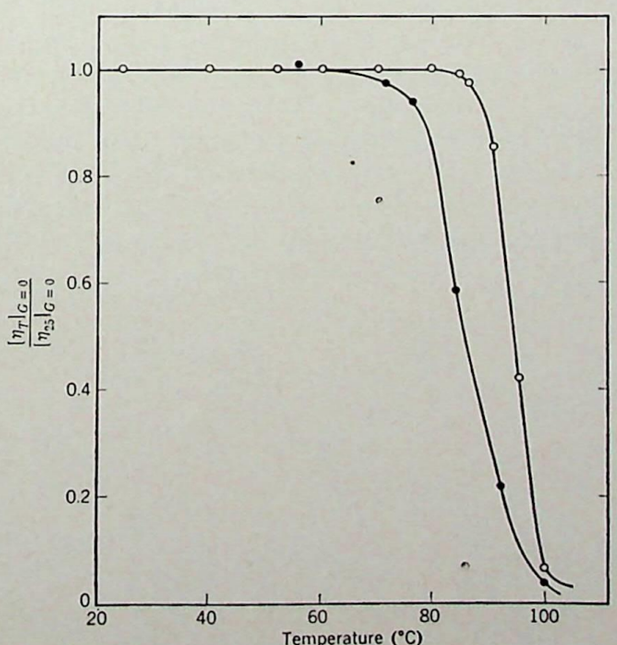


FIG. 7. The intrinsic viscosity (measured at 25°C) of DNA as a function of the temperature to which it has been heated for one hour.

Fresco, Klemperer, and the author have been concerned with self-pairing, as typically exhibited by Poly A.[38,39] We first investigated the sedimentation and intrinsic viscosity of a series of such polymers of different molecular weight at neutral *p*H. The molecular-weight dependence of these properties was found to be given by the molecular weight to the 0.45 and 0.65 power, respectively. These are self-consistent values for random coils. The absence of a significant birefringence of flow and the rise in viscosity that was observed when the ionic strength was lowered forced us to conclude that the configuration was that of randomly coiled, single chains. However, upon lowering the *p*H, we noticed a sharp transition from one type of ultraviolet adsorption to another. The titration curve showed a similar abruptness that could occur only if some cooperative transition were taking place. At *p*H's below this transition, the solutions showed very marked negative birefringence of flow. Moreover, the molecular weights were found to be greatly increased.

These observations clearly indicated that a cooperative association was taking place, upon lowering the *p*H through a critical value. The nature of this was further clarified in the following way. Neutral solutions of a single polymer were made up at a series of different concentrations. When these were acidified and the sedimentation and viscosities determined, the results showed the regular behavior that can be seen on the right-hand side of Fig. 8. It is seen that the molecular weights, or rather particle weights, span a twentyfold range and that the sedimentation constants and viscosities vary in the manner expected for homologous polymers. The weight and size increase with the concentration at which they were formed. The respective slopes are again self-consistent and their numerical values, 0.36 and 0.92, are indicative of a more-extended chain structure than was present in the neutral solutions. The results already mentioned at neutral *p*H are seen at the left of Fig. 8.

This evidence on the acid-stable form indicates fairly clearly that the polyadenylic-acid molecules have joined together in a regular structure. We were then able to show that the amino group of adenine is unavailable for reaction with formaldehyde in the acid-stable complex,

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

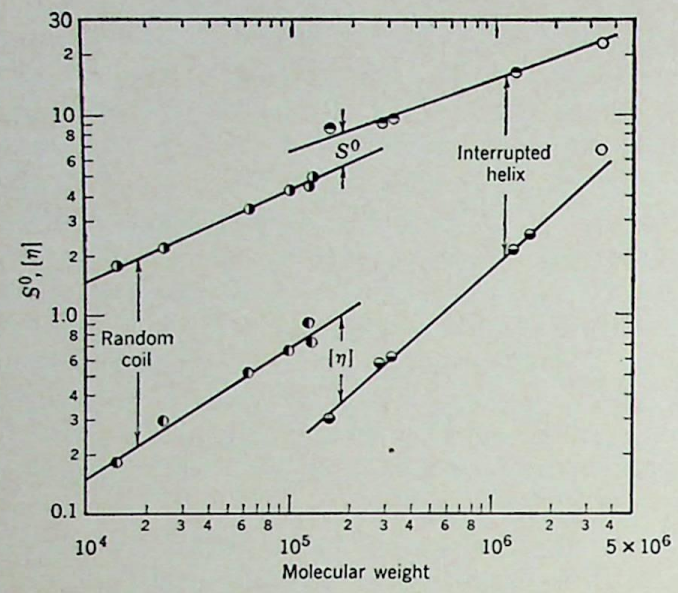


FIG. 8. The molecular-weight dependence of sedimentation constant and intrinsic viscosity for the two forms of polyadenylic acid.

but that it is highly reactive in the randomly coiled state. This showed that the pairing of the adenine groups was responsible for the association. When a model is built to satisfy this type of pairing, it is seen that it must resemble the double-helical configuration of DNA with the adenine bases nearly perpendicular to the axis. The structure requires that bases be 3.8 A apart. To test this point, J. Fresco[38] has taken x-ray photographs of solutions of Poly A (1 to 10%) above and below the transition. At low *p*H, he finds several very sharp rings, including one with a spacing of 3.8 A. At neutral *p*H, there are none. From all of this information, we conclude, therefore, that polyadenylic-acid molecules react as shown in Fig. 9. For the helix to form, the negative charge owing to the phosphate groups must be at least half-neutralized by the uptake of protons.

This behavior of Poly A is similar in many ways to other pairwise interactions that have been studied. Thus, we have nearly reached the point where all possible pairwise interactions have been cataloged. At present, it appears that, of the ten possible combinations between the four pure nucleotide polymers, 6 or 7 actually exist. The next problem is to list these pairwise interactions in order of their strength in analogy with the tables of bond energies one has in the case of covalent bonds. This information can be obtained by a careful study of the temperature at which the respective helix-coil transitions take place.

Ribonucleic Acid

Although DNA and RNA are so very similar in their chemical structure, their configurations are strikingly different. This can be said even though the configuration of RNA is only now beginning to become clear. I conclude by summarizing some investigations that have been made in this laboratory on RNA from calf-liver microsomal particles by Hall[40,41] and on RNA from tobacco mosaic virus by Boedtker.[42]

All examinations of RNA solutions have indicated that the intrinsic viscosity of such solutions is suprisingly low when compared with the molecular weight. This is quite the opposite of the case of DNA. Furthermore, RNA shows only a small birefringence of flow, and this is of the opposite sign of DNA. Thus, it is certain that the configuration of RNA in solution is not that of the extended double-stranded helix found in DNA.

An examination of the dependence of sedimentation and intrinsic viscosity on molecular weight shows that it is very close to the 0.5 power in both cases. This can be unambiguously interpreted as showing that the molecules are highly coiled and that the coils are contracted somewhat in comparison with the usual polymer chain. This result is surprising because the RNA chain carries negative charges on each repeating unit and, as a consequence, at the ionic strength (0.01 *M*) where these observations were made, it would be expected that the electrostatic repulsions would result in a very highly swollen polymer coil. If this had been the case, the intrinsic viscosities would have been much larger and the molecular-weight dependences far different. These considerations clearly suggest that there are very substantial intramolecular attractions which hold the molecules in contracted configurations.

The next problem lies in the identification of these points of internal attraction. That they indeed could be broken, and reversibly formed again, was shown in two ways.[40,41] By removing the salt from the solution, the viscosity was found to rise indicting that, by increasing the electrostatic repulsion sufficiently, the intramolecular bonds could be broken, allowing the coil to expand in the manner expected for a polyelectrolyte.

Similarly, it was found that the viscosity would rise with the temperature of the RNA solution. This again indicated the breakup of the intramolecular bonding.

It is known that RNA is hypochromic relative to its hydrolysate. Consequently, it was of interest to see if the optical density increased upon removing the salt or upon heating. Such was indeed found to be the case. A quantitative study showed that the increase in extinction was parallel with the increase in viscosity. Thus,

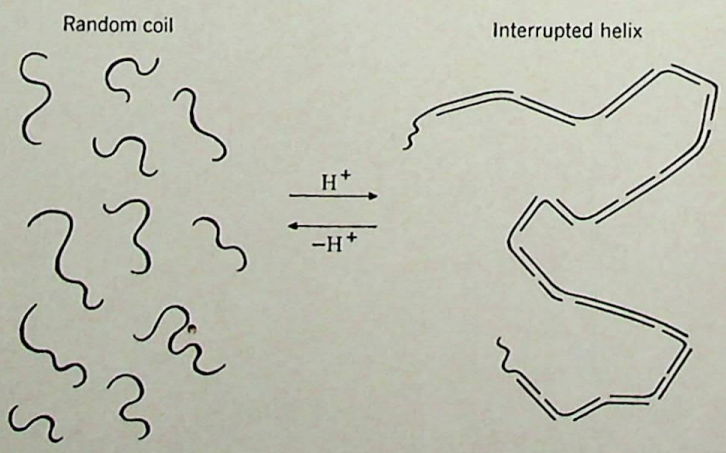


FIG. 9. Schematic illustration of the helix-coil transition in polyadenylic acid.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

the base pairing through hydrogen bonds appears to be the cause of the contracted state of the RNA.

By referring to the extinction-coefficient changes accompanying helix-coil transitions in the study of polynucleotides, it is possible to estimate the fraction of base pairs that exist in RNA solutions at moderate ionic strength and at room temperature. This comes out to be the surprisingly large figure of about 50%. At present, this conclusion can only be tentative, but it does raise a very interesting question. The studies of polynucleotide interactions reveal the large number of different types of base pairing that are possible among the bases occurring in RNA. We now have the strong indication that such pairing does occur to a great extent. Obviously, the next problem to face is whether or not such extensive base pairing occurs in a random fashion, or whether there is sufficient organization in the RNA structure to justify describing it as secondary structure.

Finally, it is important to recognize that RNA occurs in the cytoplasm, not as a discrete substance, but in intimate association with protein in the microsomal particles. Unlike DNA, which can of itself be biologically active (as revealed in bacterial transformation), RNA performs its function as a complex with protein, and it is to the analysis of the configuration of both components in these particles that much future effort is certain to be directed.

BIBLIOGRAPHY

[1] A more detailed exposition may be found in P. Doty, Proc. Natl. Acad. Sci. U. S. **42**, 791 (1956).

[2] L. Pauling, R. B. Corey, and H. R. Branson, Proc. Natl. Acad. Sci. U. S. **37**, 205 (1951).

[3] L. Pauling and R. B. Corey, Proc. Natl. Acad. Sci. U. S. **37**, 235 (1951).

[4] M. F. Perutz, Nature **167**, 1053 (1951).

[5] P. Doty, A. M. Holtzer, J. H. Bradbury, and E. R. Blout, J. Am. Chem. Soc. **76**, 4493 (1954).

[6] C. Cohen, Nature **175**, 129 (1955).

[7] P. Doty, J. H. Bradbury, and A. M. Holtzer, J. Am. Chem. Soc. **78**, 947 (1956).

[8] J. T. Yang and P. Doty, J. Am. Chem. Soc. **78**, 498 (1956).

[9] P. Doty, A. Wada, J. T. Yang, and E. R. Blout, J. Polymer Sci. **23**, 851 (1957).

[10] A. Wada and P. Doty, J. Am. Chem. Soc. (to be published).

[11] See, e.g., P. Doty and E. P. Geiduschek in *The Proteins*, H. Neurath and K. Bailey, editors (Academic Press, Inc., New York, 1953), Vol. I, p. 393.

[12] See, e.g., R. B. Simpson and W. Kauzmann, J. Am. Chem. Soc. **75**, 5139 (1953).

[13] J. T. Yang and P. Doty, J. Am. Chem. Soc. **79**, 761 (1957).

[14] J. Applequist and P. Doty, Abstracts, Am. Chem. Soc. Meeting, San Francisco (April 7–11, 1958).

[15] J. T. Yang and P. Doty (unpublished results).

[16] P. Doty, K. Imahori, and E. Klemperer, Proc. Natl. Acad. Sci. U. S. **44**, 424 (1958).

[17] W. Moffitt, J. Chem. Phys. **25**, 467 (1956).

[18] D. D. Fitts and J. G. Kirkwood, Proc. Natl. Acad. Sci. U. S. **42**, 33 (1956).

[19] D. D. Fitts and J. G. Kirkwood, J. Am. Chem. Soc. **78**, 2650 (1956).

[20] W. Moffitt and J. T. Yang, Proc. Natl. Acad. Sci. U. S. **42**, 596 (1956).

[21] W. Moffitt, D. D. Fitts, and J. G. Kirkwood, Proc. Natl. Acad. Sci. U. S. **43**, 723 (1957).

[22] D. D. Fitts and J. G. Kirkwood, Proc. Natl. Acad. Sci. U. S. **43**, 1046 (1957).

[23] K. Imahori, E. Klemperer, and P. Doty, Abstracts, Am. Chem. Soc. Meeting, Miami, Florida (April 1–7, 1957).

[24] K. Imahori, E. Klemperer, and P. Doty, "Estimate of α-helical content of proteins," J. Am. Chem. Soc. (to be published).

[25] See W. Kauzmann [Ann. Rev. Phys. Chem. **8**, 413 (1957), reference 3], for a demonstration of the existence of a square term for coupled oscillators.

[26] J. C. Mitchell, A. E. Woodward, and P. Doty, J. Am. Chem. Soc. **79**, 3955 (1957).

[27] K. Imahori, P. Doty, and E. Blout (in preparation).

[28] K. Imahori, E. Klemperer, and P. Doty (in preparation).

[29] C. M. Kay and K. Bailey, Biochem. et. Biophys. Acta (to be published).

[30] C. M. Kay, Biochem. et Biophys. Acta **27**, 469 (1958).

[31] For a review, see for example: P. Doty, J. Cellular Comp. Physiol. **49**, Suppl. 1, 27 (1957).

[32] F. H. C. Crick and J. D. Watson, Proc. Roy. Soc. (London) **A233**, 80 (1954).

[33] M. H. F. Wilkins, W. E. Seeds, A. R. Stokes, and H. R. Wilson, Nature **172**, 759 (1953).

[34] S. A. Rice and P. Doty, J. Am. Chem. Soc. **79**, 3937 (1957).

[35] J. Donohue, Proc. Natl. Acad. Sci. U. S. **42**, 60 (1956).

[36] R. C. Warner, Ann. N. Y. Acad. Sci. **69**, 314 (1957).

[37] R. Rich, *The Chemical Basis of Heredity*, W. D. McElroy and B. Glass, editors (The Johns Hopkins Press, Baltimore, Maryland, 1957), p. 557.

[38] J. R. Fresco and P. Doty, J. Am. Chem. Soc. **79**, 3928 (1957).

[39] J. R. Fresco and P. Doty, "Polynucleotides II: The x-ray diffraction patterns of solutions of the coil and helical forms of polyriboadenylic acid," J. Molecular Biol. (to be published).

[40] B. Hall and P. Doty, "The configurational properties of ribonucleic acid isolated from microsomal particles of calf liver," *Symposium on Microsomal Particles*, American Biophysical Society (Pergamon Press, London, to be published).

[41] B. Hall and P. Doty, "The preparation and physical chemical properties of ribonucleic acid from microsomal particles," J. Molecular Biol. (to be published).

[42] H. Boedtker, "Some physical properties of infective ribonucleic acid isolated from tobacco mosaic virus," Biochem. et Biophys. Acta (to be published).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

13.

Application of Infrared Spectroscopy to Biological Problems

G. B. B. M. SUTHERLAND

National Physical Laboratory, Teddington, Middlesex, England

INTRODUCTION

PROGRESS in biology at the molecular level is becoming increasingly dependent on physical and physicochemical techniques such as x-ray analysis, radioactive tracers, chromatography, light scattering, and spectroscopy. The purpose of this paper is to give a broad review of the actual and potential contributions of infrared spectroscopy to biology. Since several excellent articles and papers have already been written giving detailed accounts of the practical application of infrared methods to selected problems in biology, detail is avoided as much as possible in this presentation. In the writer's opinion, infrared methods have not been sufficiently exploited in the biological field and, since this may well be because many biologists find it hard to assess their value, it is hoped that a rather general discussion of the advantages and disadvantages of infrared analysis will be of value to biologists.

The infrared region of the spectrum is conventionally defined as extending from the limit of the visible (about 7500 A or 0.75 μ) to the present lower wavelength limit of the microwave region (about 1000 μ), but only a relatively small range of these wavelengths is of much interest to biologists, *viz.*, 2.5 to 15 μ. There are two reasons for this: the first is that within these limits are found the vast majority of spectra which throw any light on molecular structure; the second is that this region can be conveniently scanned and recorded in about ten minutes by a commercially available instrument with a rocksalt prism as the dispersing element. The exclusion of infrared wavelengths beyond 15 μ does not seem so strange when expressed in frequency* rather than wavelength units (i.e., 670 to 10 cm^{-1}), since the majority of the fundamental vibration frequencies of molecules which can be readily assigned to highly localized motions of one or both of the atoms in a chemical bond lies between 4000 cm^{-1} (2.5 μ) and 600 cm^{-1}. The limits of 670 and 4000 cm^{-1} given in the foregoing are set by the fact that beyond 670 cm^{-1} absorption by the rocksalt prism rapidly becomes complete, while above 4000 cm^{-1} the dispersion is rather low. However, it is becoming easy to replace prisms by gratings, and so these limits need no longer be set by the experimental technique. The first reason given is the dominant one, *viz.*, most of the interesting absorption bands which can be interpreted with certainty lie between 4000 and 600 cm^{-1}.

* The conventional frequency unit in infrared spectroscopy is the wave number (cm^{-1}), which expresses the number of waves in 1 cm.

The general interpretation of the infrared spectra of molecules is very straightforward; even the detailed interpretation is not difficult, provided the objective is, for instance, the identification of a compound in a mixture, or the establishment of the presence of a characteristic chemical grouping in a molecule. But, to determine the configuration of a biological polymer such as a protein is a difficult and complex problem which requires the same degree of expertise as that needed in the x-ray analysis of protein structure.

GENERAL INTERPRETATION OF INFRARED SPECTRA OF MOLECULES

The average polyatomic molecule of biological interest has an absorption spectrum in the region between 600 and 4000 cm^{-1} consisting of 20 to 30 discrete absorption bands of which about 5 to 10 will usually be much more intense than the remainder. This is because certain of the modes of vibration of such a molecule cause a periodic change in the dipole moment of the molecule. Thus, radiation of the appropriate frequency incident on the molecule can be absorbed and excite the corresponding mode of vibration. Some bands are more intense than others because the change in dipole moment is greater for some of the modes of vibration. Theoretically, a polyatomic molecule (of n atoms) has an infinite number of modes of vibration, all of which can be built up from the $3n$-6 fundamental modes of vibration. However, the change in dipole moment associated with overtone and combination frequencies is usually an order of magnitude less than that associated with the fundamental modes, and in the region of the spectrum under discussion the majority of the observed bands can be assigned to fundamental modes.

How can these fundamental modes be vizualized? Although, strictly speaking, each fundamental mode involves some motion of every atom in the molecule and would, therefore, seem to be hard to predict and describe, it turns out that an appreciable fraction of the fundamental modes of any molecule can be assigned to easily vizualized vibrations of individual chemical bonds, or of small structural units in the molecule. This is especially true of the fundamental modes which are associated with the motions of hydrogen atoms. For instance, any molecule containing an NH group will be found to have a fundamental vibration near 3300 cm^{-1} owing to the stretching and contraction of the NH bond and two other fundamentals between 1600 and 600 cm^{-1} in which the motion of the hydrogen atoms is almost perpendicular to the NH bond. More generally, in the

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

region between 3700 and 2500 cm^{-1} are found fundamental modes owing to such "hydrogenic-stretching frequencies" in the following order:

OH 3650—3150 cm^{-1},
NH 3500—3150 cm^{-1},
CH 3150—2850 cm^{-1},
SH 2650—2550 cm^{-1},
PH 2450—2300 cm^{-1}.

No other fundamental modes appear in this region of the spectrum, and so identification of any of these bonds in a compound is generally easy. There are occasional limitations, e.g., difficulty of distinguishing OH from NH, and the weakness of the SH fundamental; these are not discussed here.

The corresponding "hydrogenic deformation frequencies" for OH, NH, and CH occur between 1650 and 600 cm^{-1}, but cannot be characterized so simply. Moreover, many other fundamental modes occur within this range whose assignment and identification present special problems. However, several groups (such as CH_2 and CH_3 and the peptide link) can readily be identified by characteristic fundamentals of this type.

Before leaving the hydrogenic frequencies, mention should be made of the ease with which hydrogen bonding can be detected through its effect on these frequencies. The stretching frequencies are lowered by a few hundred cm^{-1}, and the corresponding absorption bands become quite broad and diffuse. The deformation frequencies are increased to a much smaller degree and the broadening is not always so marked.

In the region between the hydrogenic-stretching and hydrogenic-deformation frequencies (*viz.*, 2300 to 1650 cm^{-1}) occur two other separable classes of fundamentals: firstly, those owing to the stretching of triple bonds such as C≡N and C≡C occurring between 2250 and 2100 cm^{-1}, and, secondly, those largely localized in the stretching vibrations of double bonds such as C=O, C=C, C=N. The latter start about 1800 cm^{-1} and overlap a little the high-frequency end of the hydrogenic-deformation frequencies at 1650 cm^{-1}. This means that it is occasionally difficult to decide whether a band occurring between 1650 and 1500 cm^{-1} is the result of the stretching of a double bond or of the deformation of a hydrogen atom. If the hydrogen atom can be substituted by deuterium, this uncertainty is often resolved, as a hydrogenic frequency will be reduced by almost a factor of $\sqrt{2}$, whereas the double-bond frequency will be unaffected. It should be added that occasionally a hydrogenic-deformation frequency interacts with a double-bond frequency, and the two fundamentals resulting can only be described as a superposition of these two motions.

A very large number of the chemical bonds in any polyatomic molecule are of the type C—C, C—O, C—N. Such bonds do not give rise to localized fundamental vibrations. The reason is that the masses of these atoms are not very different, nor are the restoring forces between them. This means that their characteristic stretching frequencies (which can be observed in small molecules such as ethane or methyl alcohol, where they occur in isolation) are all about the same magnitude (approximately 900 to 1100 cm^{-1}). Consequently, in a polyatomic molecule, strong coupling occurs and the resulting fundamental modes involve simultaneous motions of all of the bonds of this type in the molecule. Such fundamentals are usually referred to as "skeletal modes" and have a range between about 800 and 1250 cm^{-1}. The pattern of these skeletal frequencies is often the most characteristic physical property of a molecule. Two hydrocarbons or two steroids, which may be very hard to differentiate by chemical means or by other physical properties (e.g., refractive index or melting point), can usually be recognized instantly by comparing their infrared spectra between 700 and 1300 cm^{-1}.

APPLICATION TO BIOLOGICAL PROBLEMS

There are three principal ways in which infrared spectroscopy can be of help in molecular biology, *viz.*:

(a) as an analytical tool, (b) as a means of establishing the structural formula of a biologically important compound, and (c) as a means of determining the spatial configuration of biological polymers.

(a) Analysis

Infrared spectroscopy is now a well-established and widely used analytical tool in any up-to-date biochemical laboratory. The principal advantage of infrared over visible or ultraviolet spectroscopy is that the spectrum is much richer, since it is the result of vibrations in every part of the molecule. Spectra in the visible and ultraviolet arise from the excitation of an electron in one part of the molecule, usually in a double bond such as CO group. In analytical work, one therefore has to be certain that this group is not present in any of the other compounds in the mixture. On the other hand, infrared methods will usually reveal every one of the compounds in a mixture. The only serious disadvantage of infrared analysis is that the use of water as a solvent is generally impossible because of its intense absorption over a good deal of the working range. This can frequently be overcome either by the use of other solvents, or (in the case of a solid) by making a pressed disk of the compound ground very finely in an excess of KBr. One other point should be mentioned, since it is frequently very important in much biological work, *viz.*, the minimum quantity which is necessary in order to make an identification. It is impossible to give a precise figure here, because the intrinsic intensity of absorption in the infrared varies over such a wide range between compounds, and reference must be made to the literature on this subject. As a general guide, one may say that, by using specially designed cells, it is possible to get a spectrum from about a milligram of material, but that, by using a reflecting

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

microscope attachment, this limit can be reduced to below one microgram.

(b) Structural Formulas

One of the earliest applications of infrared spectroscopy to the structure of a biologically important molecule occurred in the case of penicillin. Here, the chemists were unable to agree on which of the three possible formulas was the correct one. Although x-ray analysis gave the first and most unequivocal proof that the β-lactam formula was the correct one, infrared methods gave an independent proof and demonstrated that no structural changes took place in going from the solid state (required for x-ray work) into solution, in which penicillin is known to be a rather labile molecule. This work depended on differences in the type and in the environment of the double bonds in the three possible structures. By investigating the spectra of model compounds containing double bonds of various types in appropriate environments, a decision could finally be made in favor of the lactam structure.

The biological molecules which have been most extensively investigated by this method are the steroids. In such compounds, it is now possible to answer quite a variety of important structural questions concerning the position of CO, OH, CH_2, and C=C groups. On the other hand, it is still not possible to recognize the class of steroids by any common feature running through all of their spectra, although certain closely related groups of steroids do show common features. We find the opposite extreme in the proteins which, as a class, are easily recognized and differentiated from nucleic acids and lipids, but show remarkably little variation among one another. The reason is that the spectrum of any protein is dominated by several intense bands arising from the identical peptide links. Only in the few cases where two or three of the constituent amino acids are in great excess is it possible to identify individual amino-acid residues (e.g., glycine and alanine in silk).

The general line of attack is to compare the spectrum of the compound whose structure is unknown with the spectra of compounds of known structure containing groups of atoms identical with and in a similar environment to those suspected of occurring in the unknown. Extensive collections of infrared spectra have now been compiled by various laboratories and organizations, and the comparisons can be made relatively rapidly by various mechanized sorting devices. At first, it is usually advisable to call in an experienced spectroscopist for final confirmation, but any laboratory which takes up infrared analysis in a particular chemical field is not long in developing its own expert. The logic behind the whole process is not unlike that behind the chemical attack on the same problem. The chemist identifies certain groups in an unknown structure by their reactive properties, and he degrades the molecules into simpler ones which he can identify from their well-known chemical or physical properties. The spectroscopist identifies individual chemical bonds in an unknown structure from their known spectroscopic earmarks and tries to supplement this by identifying groups or constituent units in a structure (e.g., peptide link, benzene ring) in a similar way. The most efficient approach is, of course, to combine the two methods.

(c) Determination of Spatial Configuration of Biological Polymers

In order to understand many key biological phenomena, such as the mode of action of genes, protein synthesis, or the production of antibodies, at the molecular level, it will be necessary to establish in detail the molecular configuration of proteins and nucleic acids. This must be done by physical methods, of which the most successful so far has been x-ray analysis. However, in spite of the intensive efforts of several groups of very able workers in various countries over the past twenty years, the spatial configuration of any globular protein is still unknown. While it is true that one now seems to be very close to this goal in the case of myoglobin, the extension of the methods used for that protein to other proteins presents formidable problems. At present, infrared analysis seems to be the second most powerful physical method. Although it has not yielded such precise results, it has often given independent confirmation of the main features of a structure. More important is the fact that it can often give a lead at any early stage, either in laying down certain conditions, which the correct model structure must fulfill, or in ruling out certain models which might appear to fit a preliminary analysis of the x-ray data. Just as the x-ray method requires the polymers to be arranged in an orderly pattern (preferably in a single crystal), so the infrared method can be more precise the closer the arrangement of the polymers approaches that found in a single crystal. This is because the infrared method is based on observing the dichroism associated with key absorption bands when the spectrum is obtained using polarized radiation.

An oversimplified example will make this clearer. In the infrared spectrum of deoxyribonucleic acid is an absorption band which can be assigned to vibrations localized in the planes of the purine and pyrimidine bases. In the Watson-Crick model of DNA, these bases are nearly perpendicular to the axis of the double helix. If a highly oriented specimen of DNA is prepared in which the axes of the helices are roughly parallel to the direction of orientation, then this absorption band is very intense when the incident infrared radiation is polarized with the electric vector perpendicular to the direction of orientation, and very weak when the direction of polarization is rotated through a right angle (Fig. 1). This gives a general confirmation of one aspect of the Watson-Crick model. If it were possible to produce a single crystal of DNA of suitable dimensions in which all of the bases were parallel to each other, then the dichroism would become perfect (i.e., no absorption for

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

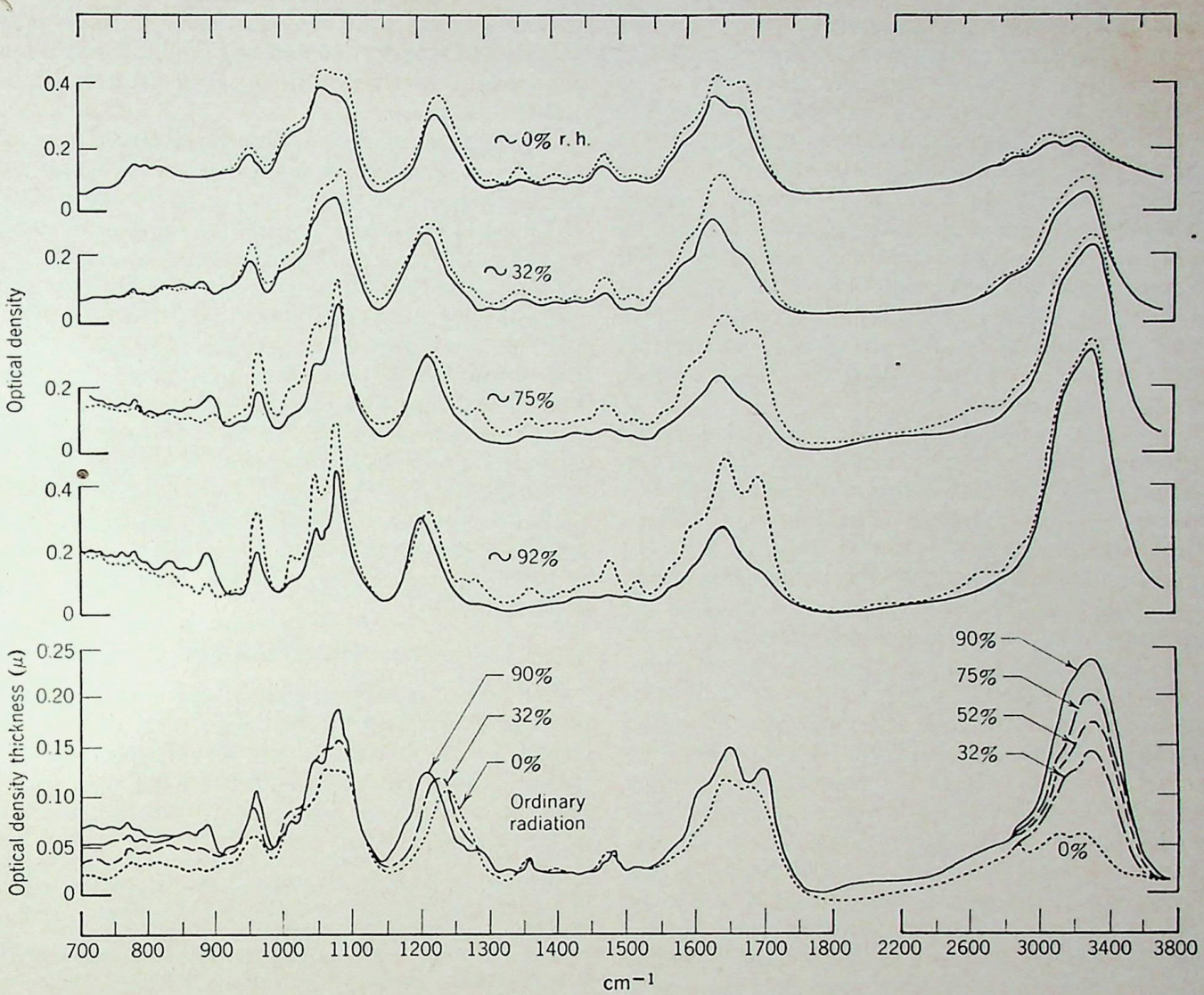


FIG. 1. The infrared absorption spectrum of NaDNA at various relative humidities. ——— indicates electric vector parallel to the orientation direction; . . . . . indicates perpendicular to the same direction. The bands referred to in the text are those between 1600 and 1700 cm^{-1} [from G. B. B. M. Sutherland and M. Tsuboi, Proc. Roy. Soc. (London) **A239**, 446 (1957)].

one polarization direction) and a precise confirmation of this aspect of the model could be given.

From the foregoing, it might appear that the infrared methods should, therefore, be applied to single crystals of biological polymers. Unfortunately, there are severe practical difficulties about such an approach. The first is that the "suitable dimensions" referred to in the foregoing hardly ever can be realized, *viz.*, a flat crystal about 10 μ thick and 2×5 mm in area. A second difficulty is that many biological polymers crystallize only in association with a high proportion of water. This limits observation to the bands which are not obscured by the spectrum of H_2O or D_2O. The third is that the molecular configuration is unknown and assumptions have, therefore, to be made about the orientations of recurring chemical groups or bonds in the polymer. The number of reasonable assumptions from which a choice has to be made can be quite large, and infrared analysis may not prove to be a sensitive enough discriminant between the various possibilities. In cases where the number of ways of fitting a long polymer into the unit cell is severely restricted, this method should give a precise answer, and it may be that the first experimental difficulty referred to can be overcome by investigating with a microspectrometer the reflection spectrum, instead of the absorption spectrum. Another line of attack is to measure the dichroism of the overtone and combination bands above 4000 cm^{-1}, since these are much weaker than the fundamentals and can be conveniently studied in a reasonably thick crystal. It should be added that there are still some theoretical problems to be overcome in the interpretation of such spectra.

Because of all of these difficulties, most of the infrared investigations to date have been made on thin sections of fibrous proteins such as porcupine quill, silk, etc. In such cases, the orientation is far from perfect and obviously the degree of disorientation must be estimated quantitatively before any reliable deductions can be made from the partial dichroism. A method of doing this has been recently worked out and applied to a number of proteins, *viz.*, porcupine quill, elephant hair, horse hair, horn, feather quill, silkworm gut, and colla-

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

gen. For each of these proteins, seven different structures were tested, and it was possible to reject all but one or two of the model structures in most cases.

It must be emphasized that this method of attack depends on certain assumptions which may not always be correct. These assumptions are as follows:

(1) The direction of change of electric moment (associated with a particular absorption band) bears the same relation to the repeat unit of the polymer as that determined from infrared studies on single crystals of small molecules which contain the repeat unit; (2) the spatial distribution of the polymer chains has a certain degree of symmetry; (3) the specimen studied consists of one structural species.

The first assumption is necessary, because it is found from work on single crystals of small molecules containing a peptide link that the change of dipole moment in a bond stretching fundamental vibration (e.g., NH or CO) is not exactly along the bond in question, but may make an angle of as much as 20° with it. Thus, the directions of these bonds (and consequently the molecular configuration) cannot be determined unless the corresponding angles are known for a peptide link in a protein. Up to the present, single crystals of only two different molecules have been investigated. The angles in these two crystals did not differ by more than five degrees. However, more work on single crystals of a few other molecules is required in order to estimate accurately the degree of uncertainty arising from this assumption. The second assumption can be shown to be a reasonable one if the sample is prepared in a certain way. At first sight, the third assumption seems reasonable and is made by the x-ray analysts. The reason for doubting it is that the structure of many of the infrared bands of proteins is complex. Whether this arises from the presence of two different configurations, or from differences in frequency between a paracrystalline and less-ordered arrangements of the polymer molecules, cannot easily be determined. In simple synthetic polymers, such as polyethylene, differences in frequency are found between crystalline and amorphous forms of the polymer. Structure can also occur in the absorption band of a single crystal because of interaction between neighboring molecules.

It appeared at one time that infrared analysis might be able to distinguish between various configurations of a protein molecule (specifically between the α-helix and the fully extended form of the polypeptide chain), even when the polypeptide chains were not arranged in any ordered pattern. This method depended on a change in the frequency of the CO absorption band in synthetic polypeptides which could be prepared in either of these two forms by precipitation from the appropriate solvent. However, there are so many possible reasons for a small change in the frequency of an absorption band that the application of any such rule to proteins (even on an empirical basis) leads to serious difficulties and inconsistencies. Until the reasons for these changes are well understood from the study of simpler model compounds, it seems very unwise to use them as a guide to protein structure.

The general situation with the infrared analysis of protein structure is similar to that which existed at an earlier stage in the x-ray analysis of the same problem. Before the interatomic distances in the peptide link were established by careful work on small molecules, it was impossible to establish the correctness of any protein structure by x-ray methods. One could only say that such and such a structure was generally consistent with the diffraction pattern. Now that these distances are known and that single crystals of proteins can be prepared containing a heavy atom, it is becoming feasible to locate the individual atoms. Similarly, before the infrared method can become precise, more work is required on single crystals of simple compounds containing the peptide link and on the experimental problems of obtaining satisfactory spectra from single crystals of proteins.

BIBLIOGRAPHY

The following bibliography is not meant to be comprehensive. It gives references to most of the general articles on the subject within which references can be found to the original papers. A few references have been given to very recent papers which are not dealt with in the general articles.

M. Beer, G. B. B. M. Sutherland, K. N. Tanner, and D. L. Wood, "Infrared spectra and structure of proteins," Proc. Roy. Soc. (London) **A249**, 147 (1959).

L. J. Bellamy, *The Infrared Spectra of Complex Molecules* (Methuen and Company, Ltd., London, 1954).

E. R. Blout, M. Parrish, Jr., G. R. Bird, and M. J. Abbate, "Infrared microspectroscopy," J. Opt. Soc. Am. **42**, 966 (1952).

C. Clark, "Infrared spectrophotometry" in *Physical Techniques in Biological Research*, G. Oster and A. W. Pollister, editors (Academic Press, Inc., New York, 1955), p. 205. (This is the best general review of the subject.)

K. Dobriner, E. R. Katzenellenbogen, and R. N. Jones, *Infrared Absorption Spectra of Steroids* (Interscience Publishers, Inc., New York, 1953).

A. Elliott, "The structure of polypeptides and proteins" in *Proceedings of the Third International Congress of Biochemistry, Brussels, 1955* (Academic Press, Inc., New York, 1956), p. 106.

R. D. B. Fraser, "Infrared spectra of biologically important molecules" in *Progress in Biophysics*, J. A. V. Butler and J. T. Randall, editors (Pergamon Press, London, 1953), Vol. III, p. 47.

Biological Applications of Infrared Spectroscopy, Ann. New York Acad. Sci., **69**, 1–254 (1957).

H. M. Randall, R. G. Fowler, N. Fuson, and J. R. Dangl, *Infrared Determination of Organic Structures* (D. Van Nostrand Company, Inc., Princeton, New Jersey, 1949).

G. B. B. M. Sutherland, "Infrared analysis of the structure of proteins and nucleic acids" in *Convegno Antonio Baselli le Macromolecole Viventi*, Ist. Lombardo sci. lettere **89**, 67 (1955).

G. B. B. M. Sutherland and M. Tsuboi, "Infrared spectrum and molecular configuration of sodium deoxyribonucleate," Proc. Roy. Soc. (London) **A239**, 446 (1957).

D. L. Wood, "Infrared spectrophotometry" in *Methods in Enzymology* (Academic Press, Inc., New York, 1957), Vol. IV, p. 104.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

14
Concentrated Macromolecular Solutions

BRUNO H. ZIMM

Physical Chemistry Section, General Electric Research Laboratory, Schenectady, New York

DILUTE solutions were covered in the papers of Rice (p. 69), Stockmayer (p. 103), and Doty (p. 61). This paper first discusses concentrated solutions in contrast to dilute solutions. Many properties in concentrated solutions are determined by the fact that the interactions between the molecules are of primary importance, whereas they are of secondary importance when the dilution is high.

In most macromolecular solutions, one may never neglect completely the interactions between the large molecules. The nearest exception to this is probably solutions of the globular proteins at high ionic strength and at reasonable concentrations—i.e., concentrations in which it is quite easy to do physical experiments, such as the determination of intrinsic viscosity, and so on. In any solution, interactions can be ignored fairly well, providing highly precise measurements are not desired. As soon as extended particles—like fibrinogen, or especially the nucleic acids, or the randomly coiled polymers such as the synthetic ones, or starch or other polysaccharides—become involved, it is practically impossible to make them sufficiently dilute to eliminate the interactions between the molecules, and, at the same time, to have enough material remaining in solution to measure by the standard techniques. In these cases, it might be said that one is dealing always with concentrated solutions.

For practical purposes, the interaction can be characterized by the following. The osmotic pressure *versus* the concentration is plotted. The ideal-solution laws predict that the osmotic pressure divided by the concentration should be practically constant. The van't Hoff law says that the osmotic pressure is proportional to the concentration. For an ideal solution, one obtains, then, the curve *a* in Fig. 1. Most of the time, however, curve *b* is obtained. As soon as the curve has deviated from the ideal law by about as much as the ideal law itself predicts, one has a concentrated solution.

This upward, or positive, deviation usually results from repulsion between the molecules, and the fact that the molecules are large means that there are unavoidably a lot of repulsions just from their space-filling properties. This is the common type of behavior in a macromolecular solution; in effect, the molecules elbow each other apart. There is the same type of behavior in the theory of imperfect gas where one must consider the "co-volume" of the molecules. The pressure is raised in consequence.

A similar effect is observed on the scattering of light. If there are these repulsive interactions, the fluctuations of concentrations can not take place as easily or as independently of each other, and with less fluctuation there is less scattering of light.

These are manifestations of the intermolecular forces. There is one trap into which the uninitiated easily fall, and that is to consider these intermolecular forces as acting as though the molecules were in a vacuum. The molecules are in a solvent, and the forces between the molecules and the solvent molecules and other solvent molecules are as important as the direct interaction between the macromolecules. What is seen here actually is a balance of forces.

Figure 2 shows some data obtained by Doty and Yang[1] on polybenzylglutamate. The fraction of internal hydrogen bonds in a molecule are shown. When they are all intact, the helical form results, and when none of them are intact, the random-coil form results. The transition region between these two forms depends upon the solvent composition, and this particular solvent is a mixture of dichloroacetic acid and dichloroethane. Taking the transition temperature as the reference point, it is seen that, when the temperature is reduced, one goes to the unbonded form. If the temperature is raised, one goes to the bonded form. It is incorrect to consider solely the interactions between parts of the macromolecule itself because, in that case, an increase in temperature should melt it— that is, break the bonds. In fact, just the opposite is done in raising the temperature; the bonds are formed. Obviously, there is competition between the bonding to the solvent and the internal bonding of the macromolecule. The solvent here actually is the determining factor. In fact, the entropy of this transition has the wrong sign. Instead of the entropy of the random coil being the greater and that of the helix the lesser, the reverse is true. The helical

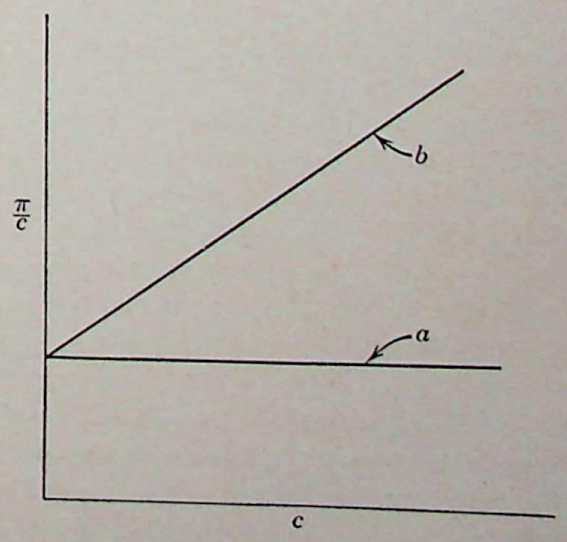


FIG. 1. Osmotic pressure π as a function of concentration c for real and ideal macromolecular solutions. See text.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

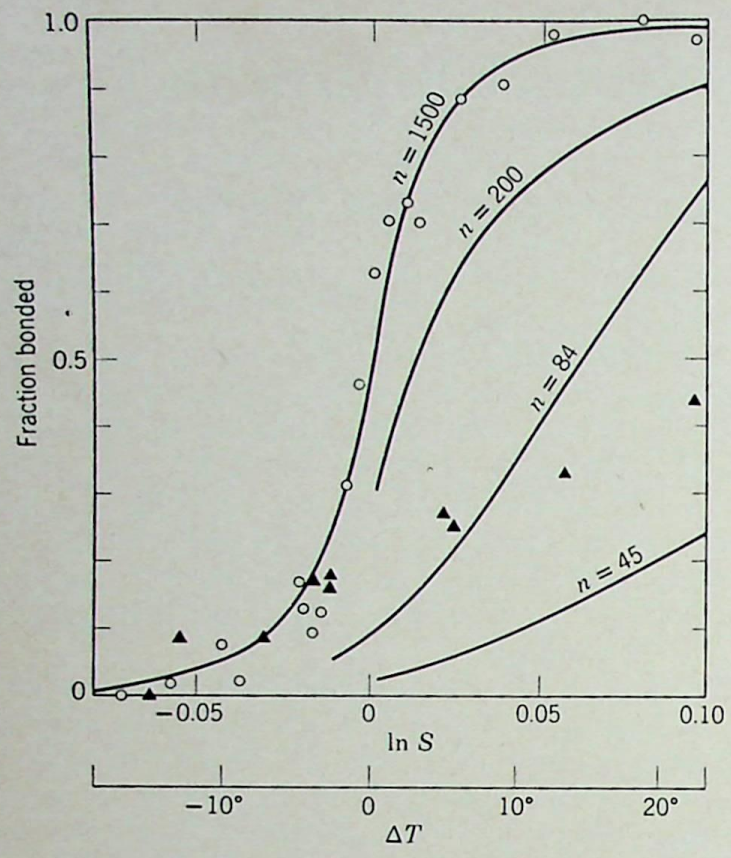


FIG. 2. Fraction of intramolecular hydrogen bonds in poly-γ-benzyl-L-glutamate; points calculated from optical-rotation data of Doty and Yang[1]; the curves represent an unpublished theory developed by J. K. Bragg and the author.

form is the one that is stable at high temperatures and the entropy of the helix is greater and that of the random coil is less. This means that the solvent is participating and that there is considerable entropy change in the solvent. This illustrates the pitfalls possible when the solvent is overlooked.

Consider highly concentrated solutions, such as 10 to 50% (or possibly more) dissolved material in solvent. What can be said about them? The kinetics of such systems are treated in the paper by Ferry (p. 130). The concern here is with statics. One of the first items of interest is, "What is the absorptive capacity of such solutions?" They can be called gels, for convenience. What then is the absorptive capacity of gels for foreign solvents, such as ions, etc.? For gels composed of long-chain macromolecules, the thermodynamic behavior with respect to the absorption of solvent is described by a theory advanced by Flory, Huggins, and Guggenheim.[2,3] The derivation contains numerous obviously wrong assumptions, over-simplifications, and so on; however, since it is a classical theory, one need say little about it here. If one simply wants to calculate an absorption isotherm, it is useful. It has an adjustable parameter that allows one to make it work fairly well in many different cases.

Another point of view should be mentioned. As shown in Fig. 3, one can plot the volume fraction of a solute ϕ_1 as a function of the activity of the solvent a_1, defined in such a way that, in a pure solvent, a_1 has the value of unity. The activity a_1 is really little more than the relative vapor pressure—the vapor pressure of the solvent in the mixture over that of the solvent. To be very careful, gas-law corrections and such things must be included.

An ideal solution should give the perfectly straight diagonal of curve *c*. This shows the activity of the small molecule component proportional to the volume fraction. Raoult's law actually was stated in terms of mole fractions, but if it is modified slightly, it states that the ideal solution follows curve *c*. However, a real macromolecular solution usually gives something of the type of curve *d*. Sometimes the curve is more complicated, as is shown later.

Two characteristics are to be noted. At the upper right-hand corner, the slope of curve *d* is, as is usual, very much smaller than the ideal one. This is merely a manifestation of the repulsive interaction mentioned before, i.e., a positive deviation. The osmotic pressure is higher; the vapor pressure, therefore, also is higher in dilute solution than in the ideal case. At the lower left-hand corner, curve *d* usually rises much more steeply than the ideal. This slope, in cases where there are no specific interactions, is of the order of three to five times that of the ideal.

There is a very simple interpretation for this marked steepness in slope. Malcolm Dole probably was the first to apply it to any extent.[4] It is based upon the proposed mechanisms for rigid absorbing systems, like clays, inorganic catalysts, carbonblack, etc. It says that, in the absorbing material—in this case, the dry polymer—there are a number of sites which can take up the solvent, and that the rate with which the activity rises is an inverse function of the number of sites. If there is a large number of sites, then much absorbed material can be taken up without raising the activity very much. If there are a very few sites, however, they will fill up very fast, and the activity will rise much more rapidly than the volume fraction.

A mixture of two pure liquids, toluene and benzene, gives the straight line, curve *c* in Fig. 3. The number of

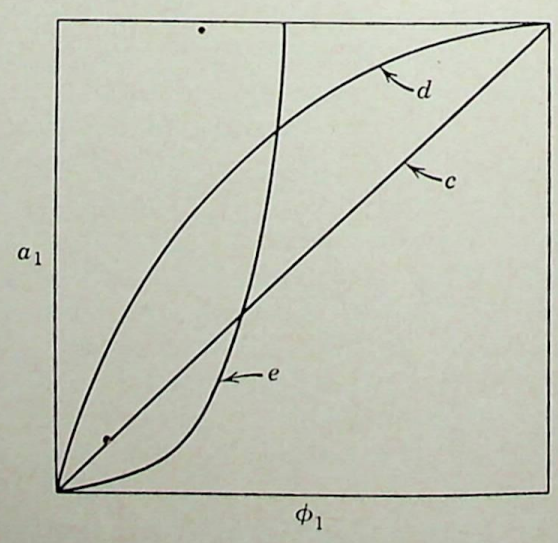


FIG. 3. Activity a_1 against volume fraction ϕ_1 of solvent sorbed in a high polymer. See text.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

sites in the mixture apparently increases in proportion to the amount of toluene put in, with one site available for every molecule added. The whole volume remains accessible to these molecules and curve *c* results.

The macromolecular system is different. There are not nearly as many available sites in which to put solvent molecules; there appear to be about one-fourth as many. This suggests a very simple and reasonable hypothesis. Figure 4(a) shows the absorption of toluene in benzene. The benzene molecules can be elbowed out of the way, presenting little resistance to the added toluene, since the whole volume is accessible to its admission.

For macromolecules, the volume initially consists of intertwined chains [Fig. 4(b)]. If a toluene molecule is introduced at position *a*, a long cavity will have to be made with considerable wasted space at its ends, since the macromolecular chains are quite rigid. Therefore, the toluene molecule enters position *b* preferentially where it can cause with greater ease the necessary rearrangements in the neighboring macromolecules. It appears obvious that material of this kind is relatively unreceptive to small molecules; it can fit them into certain specific places only.

One cannot predict the number of specific places, but the experiments quoted in the following indicate that about one-fourth of the volume is occupied by sites into which a small molecule easily can be introduced. This discussion refers to unactivated absorption where no strong specific forces are involved. An example is toluene in polystyrene where the forces are fairly well in balance.

In systems permitting hydrogen bonding, specific interactions are likely, with pronounced deviations resulting (curve *e*, Fig. 3). This curve, which is typical of water absorbed in protein, starts off with a very low, rather than a high, slope, indicating that some of the sites are very receptive to the water molecules. After these specific places have been filled up, the curve rises rapidly. It even may come up and hit the axis, which means that the protein is saturated and cannot take up any more water.

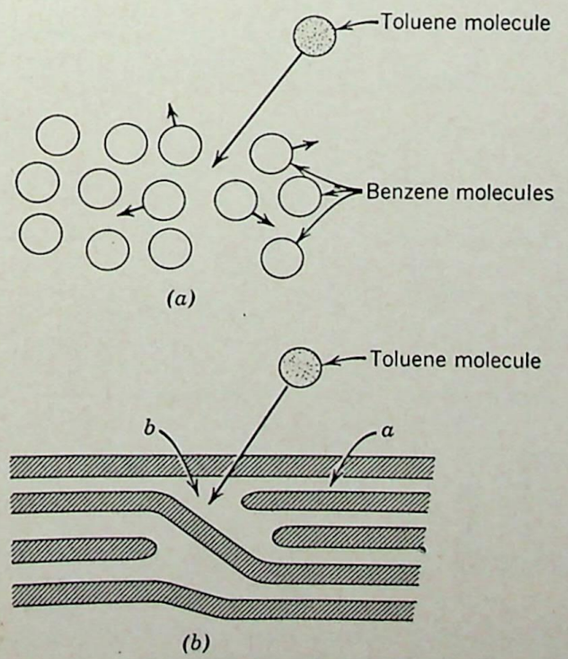


FIG. 4. Sorption of a small molecule, toluene, (*a*) in an ordinary liquid and (*b*) in a high polymer.

In this case, there are a few very specific active sites as compared with the previous case in which there are many inactive sites. But, when one attempts to apply the simple site hypothesis quantitatively, it is not very successful, because the manner in which the absorbed molecules affect the protein has been ignored. Apparently, the protein rearranges its state considerably, and it is difficult to describe the situation in simple terms. The whole concept of an inert absorptive site loses its applicability.

There is, however, another approach to this problem.[5] Define a function that measures the tendency of the absorbing molecules to cluster. This function, the "clustering function," has an exact molecular definition. It is given by an integral,

$$G_{11}=(1/V)\int\int[F_2(i,j)-1]d(i)d(j),$$

where i and j are molecules of component 1, the solvent; V is the total volume; and $F_2(i,j)$, the distribution function, is defined by the statement that $(1/V^2)F_2(i,j)d(i)d(j)$ is the probability that the molecules i and j are each at the positions specified by the coordinates (i,j) in the range of these coordinates $d(i)$ and $d(j)$. This distribution function is familiar in the case of spherical molecules where it is called the radial distribution function. It is simply the probability of finding one molecule in a given position with respect to another.

The quantity $\phi_1 G_{11}/v_1$, where v_1 is the molar volume of the solvent, is the mean number of solvent molecules *in excess of the random expectation* in the neighborhood of a given solvent molecule. Thus, it measures the clustering tendency of solvent molecules, and, for this reason, G_{11} is called the clustering function. If G_{11} is positive, the solvent molecules cluster; if it is negative, as frequently happens, they do not.

It should be noted that this clustering function has a simple and exact relation to the isotherm. There are no assumptions except for a very minor approximation of neglecting the compressibility of the system. The relation is as follows:

$$\frac{G_{11}}{v_1}=(\phi_1-1)\frac{\partial}{\partial a_1}\left(\frac{a_1}{\phi_1}\right)-1.$$

Similar relations were known to Willard Gibbs,[6] but they were neglected until interest was revived recently by Mayer[7] and by Kirkwood[8] and their collaborators.

The clustering function can be determined from the isotherms, and they immediately reveal whether the

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

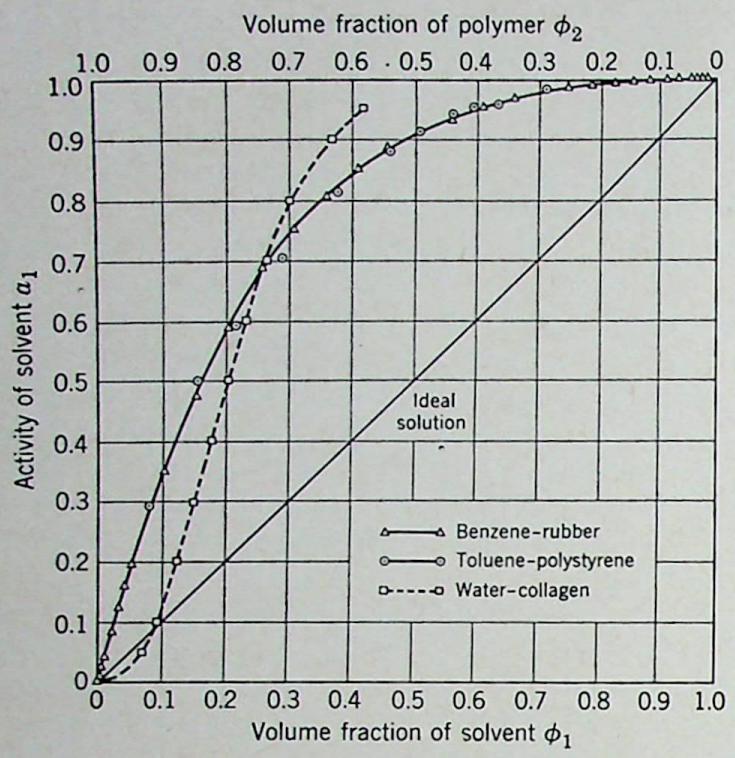


FIG. 5. Activities of solvents *vs* volume fractions of solvents and polymers for benzene-rubber, toluene-polystyrene, and water-collagen at 25° [from B. H. Zimm and J. L. Lundberg, J. Phys. Chem. **60**, 425 (1956)].

absorbing molecules tend to crowd into a few sites, or whether they tend to occupy isolated sites and prevent other molecules from entering. This would depend upon whether G_{11} is positive or negative, and upon how much so.

Figure 5 shows the data for three systems, two of which are simple nonpolar systems. The heavy solid curve represents the activity of the solvent in the systems benzene-rubber and toluene-polystyrene, in which the intermolecular forces are neither specific nor strong. The two curves lie almost on top of one another. The initial slope is quite high, which indicates that the number of sites available for the absorption of the incoming solvent molecules is much smaller than it would be in a mixture of two ordinary liquids. A water-collagen system exhibits another type of behavior.[9] First, there is a region of very strong attraction for the first few molecules of water absorbed, with the water hardly showing any vapor pressure. Then, after the initial sites are filled, the activity rises very rapidly until it is seen to come up to that of liquid water. Dole and McLaren[9] found that the first region has a strong negative heat of absorption; that is, heat was evolved to the amount of about 6 kcal/mole of water in the region of ϕ_1 less than 0.1.

Figure 6 shows the clustering function for these same systems. A clustering function of zero would mean no interactions at all between the molecules. The ideal solution has G_{11} equal to -1. In the ideal solution, there are interactions; the molecules still have volume. One molecule cannot be placed on top of another, so the clustering function is -1; in other words, the first molecule excludes one molecular volume to the other molecules. For the hydrocarbons, a positive clustering tendency appears. G_{11} is fairly uniformly $+1$ for all values of ϕ_1. This means that the molecules of the liquid tend to cluster in the polymer, which is probably a manifestation of the same geometrical problems described in Fig. 4(b). It is not easy for the polymer to make room for the liquid molecules, but when one has been admitted and a cavity is opened, it is apparently easier to put in other liquid molecules by extending the same cavity than by opening up new cavities. Therefore, there is this mild clustering tendency which is not a manifestation of any special attractive properties between the benzene and rubber or between the toluene and polystyrene. Actually, in the former case, the heat of absorption is mildly negative, and, in the latter case, it is mildly positive. This contrast leads to the tentative conclusion that G_{11} is affected more by the geometry of the packing than by the heat of sorption.

The Flory-Huggins-Guggenheim theory predicts values of G_{11} of less than unity for the systems previously referred to. In general, it underestimates the amount of clustering. In other words, it does not take sufficiently into account the intrinsic heterogeneity of the mixture. This, in fact, is very difficult to do.

The water-collagen system presents a different picture. Initially, the clustering function is markedly negative. The first experimental point seems to be at about -8, and the curve is very steep. This means that the first molecule of water that is adsorbed excludes at least eight times its own volume to other molecules. One can interpret this as meaning that the first molecule occupies a special site, and that there are no other special sites

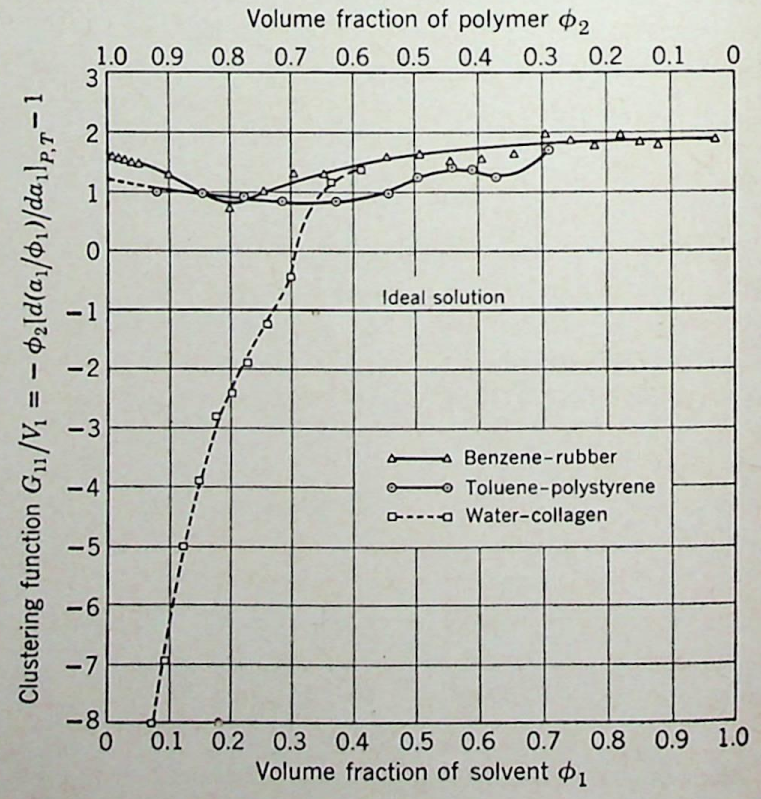


FIG. 6. Clustering functions of solvents *vs* volume fractions of solvents and polymers for benzene-rubber, toluene-polystyrene, and water-collagen at 25° [from B. H. Zimm and J. L. Lundberg, J. Phys. Chem. **60**, 425 (1956)].

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

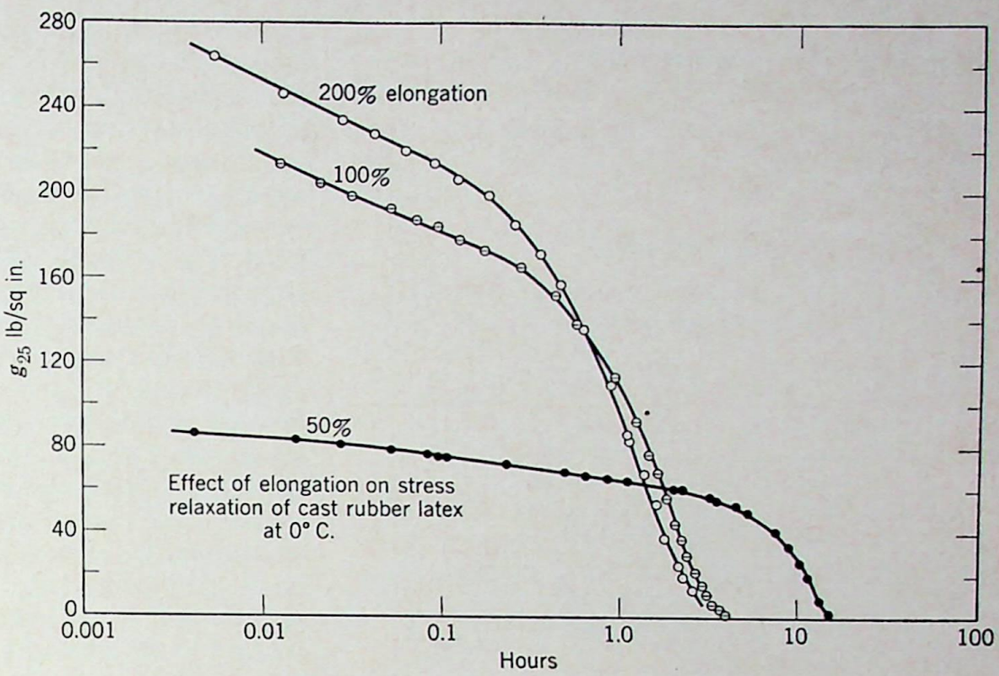


FIG. 7. Crystallization of stretched rubber [from A. V. Tobolsky and G. M. Brown, J. Polymer Sci. 17, 547 (1955)].

within at least the cube root of eight molecular diameters around. Other molecules that enter will have to find a site elsewhere. After these special sites are filled, the system behaves more normally, and just before the collagen is saturated with water, it acts like an ordinary unspecific high polymer in which the water molecules show a mild tendency to cluster.

From the curve of the clustering function, it is apparent that the initial excluded volume for water on collagen is at least 10 water volumes, or 180 ml, which amounts to less than 0.8 mole of absorption sites per 100 g of collagen. By use of the Langmuir isotherm, Dole and McLaren estimated about 0.6 mole,[9] which seems to be reasonable agreement in view of the fact that the curve is still climbing steeply at the lowest point. If the Langmuir isotherm is used, one makes special assumptions, whereas, with the clustering function, none are made.

Consider now the elasticity of macromolecular systems. Most concentrated macromolecular solutions are not simple liquids. They show some kind of structural elasticity or structural viscosity. The following discussion is restricted to the equilibrium manifestations. [For further discussion, see Ferry (p. 130).] Rubber is classically the best-known example. When a rubber band is stretched, it warms up; the work of stretching is converted into heat. In contrast, when a steel spring is stretched, there is practically no thermal effect; in steel, the work of stretching is converted into internal energy. The obvious explanation is that in rubber the coiled macromolecular chains are straightened and their entropy is reduced. If this is done reversibly, an output of heat must result. Alternatively, the retractive force of the stretched rubber band is caused by the thermal vibration of the macromolecular chains which try to return to a state of increased entropy.

This is the entropic theory of rubber elasticity, and it works very well at elongations of less than 200 to 300%. At small elongations, the change of entropy accounts for nearly all of the retractive force. This is discussed by Flory.[2]

Beyond about 300% elongation, the material stiffens rapidly and the internal energy contributes heavily to the retractive force. The energy of extension becomes negative, which means that there is a decrease in internal energy on extension. With further extension, the entropy also becomes more negative than theory predicts. This marked decrease in entropy and energy is suggestive of the decreases that occur during crystallization where the molecules are ordered with the result that the entropy decreases; at the same time, they are placed in energetically favorable configurations so that the energy decreases also.

Direct evidence for this is shown in Fig. 7 from work by Tobolsky and Brown.[10] It is known that rubber does crystallize, since x-ray patterns are obtainable. If rubber is stretched at 0°C, which is well below the freezing point for the initially amorphous rubber, the retractive force decreases with time; the curve approaches the axis. (The apparatus is incapable of measuring a negative retractive force. In many cases, the rubber band extends spontaneously.) The macromolecular chains, which initially are lined up somewhat by the stretching, are crystallizing now into an even better linear array, and the band extends.

At present, polyethylene thread is available which, when stretched, oriented, and then cross-linked by irradiation, exhibits a striking phenomenon. If heated, it melts and contracts about 10 to 20%. When cooled and frozen at about 125°C, it extends again spontaneously. This phenomenon is the result of crystallization of oriented chains. A similar process occurs in biological

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

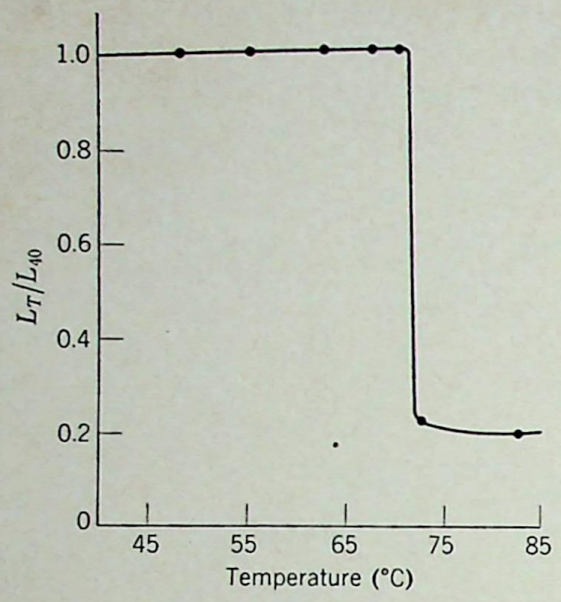


FIG. 8. Length as a function of temperature for formaldehyde-tanned rat-tail tendon under a constant small load [from P. J. Flory, J. Cellular Comp. Physiol. **49**, Suppl. 1, 175 (1957)].

materials, for example in collagen. Figure 8 illustrates some interesting data obtained by Dumitru, and discussed by Flory.[11] Collagen, normally, is in crystalline form, but, if heated to its melting point, it suddenly contracts. Unfortunately, when annealed below its melting temperature, collagen does not re-extend very markedly on cooling, although the density does go up somewhat. This striking change of length accompanying the melting of the oriented crystal is similar to the phenomenon observed in rubber.

A similar interesting calculation can be made for α-helices. Instead of plotting the usual stress-strain diagram with length as a function of force, two convenient dimensionless variables can be introduced. In place of length L, substitute $L/n\delta$, where n is the number of residues in the chain and δ is the length along the helix axis of one residue. Similarly, in place of force f, substitute $fb^2/3\delta kT$, where b^2 is the square of the effective length of a residue in the random form; that is, b^2 is the mean-square end-to-end length divided by n. These variables are plotted in Fig. 9.

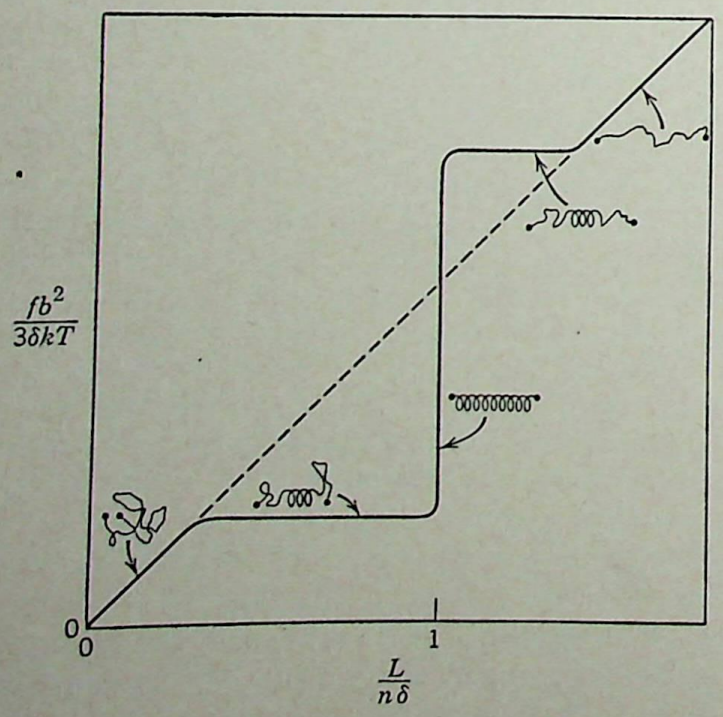
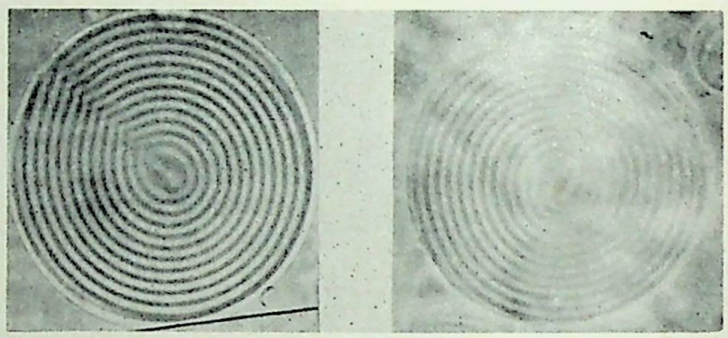


FIG. 9. Hypothetical stress-strain diagram for a polypeptide chain that can "crystallize" into a helix on stretching. See text for symbols.

Suppose that the material initially is in the random form. The stress-strain curve is linear and somewhat simpler than the curve for rubber, because only one molecule is considered rather than an array of molecules with different orientations. Initially, the slope of the curve is unity. Upon stretching, the helical form is favored because the helix is oriented in the direction of stretch. The coil-to-helix transition commences with an essentially two-phase situation. In this situation, considerable change in length results from a small change in force, since the proportion of helical to random form rapidly increases. Upon complete conversion into an α-helix, the structure becomes very rigid, and large changes in the force produce little change in length. As



(a) (b)

FIG. 10. Two spherulites of an anisotropic solution. (a) Large regular spherulite, showing radial fault. *R*1 in methylene chloride; spacing about 0.015 mm×154. (b) Large regular spherulite, fault not visible. Phase contrast; *R*1 in methylene chloride; spacing about 0.015 mm×131 [from C. Robinson, Trans. Faraday Soc. **52**, 571 (1956)].

illustrated in Fig. 9, the 45° line is crossed, and the random form represented by the dashed diagonal line would extend more readily than the helical form in the direction of stretching. Thus, the random form again becomes more favorable, and a second transition from the helical to the random coil occurs. As shown in the diagram, both of the horizontal regions of the curve (the transition regions) represent a mixture, with a segment of random coil followed by one of helix. There is probably always a random segment at the end, because of the disordering effect of the end.

An additional subject of interest is that of liquid crystals. Many macromolecular solutions contain a considerable amount of order. Tobacco mosaic virus particles are random in dilute solution, but, in a more concentrated form, the solution is intrinsically anisotropic and exhibits colors between crossed polaroids. Its x-ray diffraction pattern consists of arcs of circles. This pattern originally was investigated by Bernal and Fankuchen.[12] There are also many mixtures of soaps

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

and water and of various lipids in water which have similar properties. There is, in fact, a large and rather neglected field of what are generally called liquid crystals, or crystalline liquids, or *mesophases*—different names for the same phenomenon.

Figure 10, a recent photograph taken by Conmar Robinson,[13] shows a liquid crystal, a drop of a solution of polybenzylglutamate in methylene chloride. The concentration is roughly 20% by weight. In this solvent, the molecules are in the α-helix form, and the presence of the long extended structures promotes formation of liquid crystals. In a concentrated solution, the liquid crystals separate and form small drops of a separate phase. Figure 10 illustrates one of the drops surrounded by the ordinary isotropic solution. The diameter of the drop is approximately 300 μ. Each illustrated layer is about 10 to 15 μ thick. This onion-like structure is remarkable. In Fig. 10(a), one sees a single spiral with something of the nature of a stem, or radial fault, on one side. Looked at from a different angle, Fig. 10(b), the stem radius is not apparent; instead, a double spiral is seen. The structure is very striking. For further reference, the reader is referred to Robinson[13] and to papers by Robinson and Frank.[14] Onsager[15] and Flory[16] have proposed quantitative theories which try to explain the formations of the liquid crystal as resulting from the anisotropic interaction between long rod-like molecules. These theories indicate very nicely the essential feature that, if too many long rods are packed in a given space, they pack very poorly unless they are aligned; if aligned, they can be packed more economically. Hence, a concentrated solution of rod-like molecules tends to form liquid crystal structures, but, if diluted, isotropic structures occur. In fact, phase boundaries can result between two solutions of the same materials at different concentrations.

BIBLIOGRAPHY

[1] P. Doty and J. T. Yang, J. Am. Chem. Soc. **78**, 498 (1956).

[2] P. J. Flory, *Principles of Polymer Chemistry* (Cornell University Press, Ithaca, 1953).

[3] E. A. Guggenheim, *Mixtures* (Oxford University Press, London, 1952).

[4] M. Dole, Ann. N. Y. Acad. Sci. **51**, 705 (1949).

[5] B. H. Zimm and J. L. Lundberg, J. Phys. Chem. **60**, 425 (1956).

[6] J. W. Gibbs, *The Collected Works of J. Willard Gibbs* (Longmans, Green and Company, New York, 1931) II, 201.

[7] W. G. McMillan and J. E. Mayer, J. Chem. Phys. **13**, 276 (1945).

[8] J. G. Kirkwood and F. P. Buff, J. Chem. Phys. **19**, 774 (1951).

[9] M. Dole and A. D. McLaren, J. Am. Chem. Soc. **69**, 651 (1947).

[10] A. V. Tobolsky and G. M. Brown, J. Polymer Sci. **17**, 547 (1955).

[11] P. J. Flory, J. Cellular Comp. Physiol. **49**, Suppl. 1, 175 (1957).

[12] J. D. Bernal and I. Fankuchen, J. Gen. Physiol. **25**, 111, 147 (1941).

[13] C. Robinson, Trans. Faraday Soc. **52**, 571 (1956).

[14] C. Robinson and F. C. Frank, Discussions Faraday Soc. (to be published).

[15] L. Onsager, Ann. N. Y. Acad. Sci. **51**, 627 (1949).

[16] P. J. Flory, Proc. Roy. Soc. (London) **A234**, 73 (1956).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

15
Rheology of Macromolecular Systems

JOHN D. FERRY

Department of Chemistry, University of Wisconsin, Madison 6, Wisconsin

I. INTRODUCTION

As the paper by Zimm (p. 123) shows, mechanical energy can be stored in a flexible macromolecular network at elastic equilibrium because of the configurational restrictions imposed on it by an external deformation. The resulting rubber-like elasticity and its characteristic dependence on temperature and cross-linking density are familiar and intuitively reasonable phenomena. Less familiar is the fact that elastic energy can be stored in a macromolecular system *without* cross-links when it is subjected to steady-state flow; this elasticity is manifested in solutions of either flexible or rigid molecules and even at high dilution. In flow, of course, mechanical energy is also continuously dissipated as heat. In some cases, the energy storage is masked by the dissipation and is difficult to detect; in others, including numerous systems of biological interest, the stored energy is obvious as seen, for example, in elastic recoil after cessation of flow.

The presence of stored elastic energy is usually associated with non-Newtonian character of the flow; the apparent steady-flow viscosity falls with increasing shear rate. Moreover, the elastic strain (i.e., that part of the deformation which is recoverable after removal of stress) may not be directly proportional to shear stress, so that the elasticity may be non-Hookean. The province of rheology should include the description and explanation of these nonlinear relationships, as well as the simultaneous appearance of viscosity and elasticity. Nevertheless, for sufficiently small stresses, the deviations from Newtonian flow and Hookean elasticity can be made negligible, and most of the present discussion is restricted to such conditions, corresponding to so-called linear viscoelastic behavior. Although nonlinear phenomena are certainly important and potentially valuable sources of information, they are less well understood and are omitted for the sake of brevity.

With this simplification, a rheological description of the concomitant viscous and elastic behavior of a macromolecular system involves the following quantities:

(a) The steady-flow viscosity, $\eta = \mathfrak{T}/\dot{\gamma}$, where $\mathfrak{T}$ is the shear stress and $\dot{\gamma}$ the rate of shear. The energy dissipated per cc per sec in flow is $\eta\dot{\gamma}^2$.

(b) The steady-state compliance, $J = \gamma_e/\mathfrak{T}$, where γ_e is the elastic strain. The energy stored per cc is $J\mathfrak{T}^2/2$.

(c) Various parameters representing characteristic times. For example, the ratio of energy stored to energy dissipated per second has the dimensions of time ($=J\eta/2$). The terminal relaxation time for a macromolecular solution, τ_1, is usually also of the order of $J\eta$. The time required to attain steady-state flow under constant stress is approximately τ_1; after cessation of flow, the stored energy is dissipated by relaxation processes within an interval which is also of the order of τ_1. For a complete description of the approach to steady-state flow and of the course of relaxation, as well as of other time-dependent rheological processes, a spectrum of relaxation times is usually required.

From experimental measurements of stored energy and of the rate of approach to steady-state flow, or from equivalent information about response to oscillating stresses, conclusions can be drawn regarding the ease of motion of macromolecular segments through their surroundings and the nature of long-range coupling between molecules, as illustrated by the examples which follow.

II. ENERGY STORAGE IN STEADY-STATE FLOW

The storage of energy in very dilute solutions during flow can be attributed qualitatively to the departure from random orientation, familiar from discussions of streaming birefringence, which leads to a decrease in the entropy content in the steady state. The complete rheological behavior can be predicted by considering the interactions between the frictional forces imposed on solute molecules by the flowing solvent and their own Brownian motions, as calculated by Kirkwood and Auer[1] for thin rigid rods and by Rouse[2] and Zimm[3] for flexible coils. The results correspond mathematically to the mechanical models in Fig. 1, where the solvent viscosity η_s is shown added to various viscosity contributions from the solute. For rods, one of the solute contributions behaves as though in series with an elastic spring; for coils, there is an infinite series of solute contributions, each in series with a spring. The magnitudes of some of the model components are given in Table I in terms of the solution viscosity, η, the molecu-

TABLE I. Rheological characteristics of macromolecular solutions.

	Thin rigid rods (Kirkwood and Auer[a])	Flexible free-draining coils (Rouse[b])
G_1	$3cRT/5M$	cRT/M
η_P	$(\eta-\eta_s)/4$	...
η_1	$3(\eta-\eta_s)/4$	$6(\eta-\eta_s)/\pi^2$
J (very dilute)	$3M[\eta]^2c/20RT$	$2M[\eta]^2c/5RT$
J (concentrated)	[c]	$2M/5cRT$

[a] Reference 1.
[b] Reference 2.
[c] Rigid rods will not form a concentrated disordered solution, preferring thermodynamically to distribute themselves between a dilute disordered and a very concentrated ordered phase.[4,5]

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

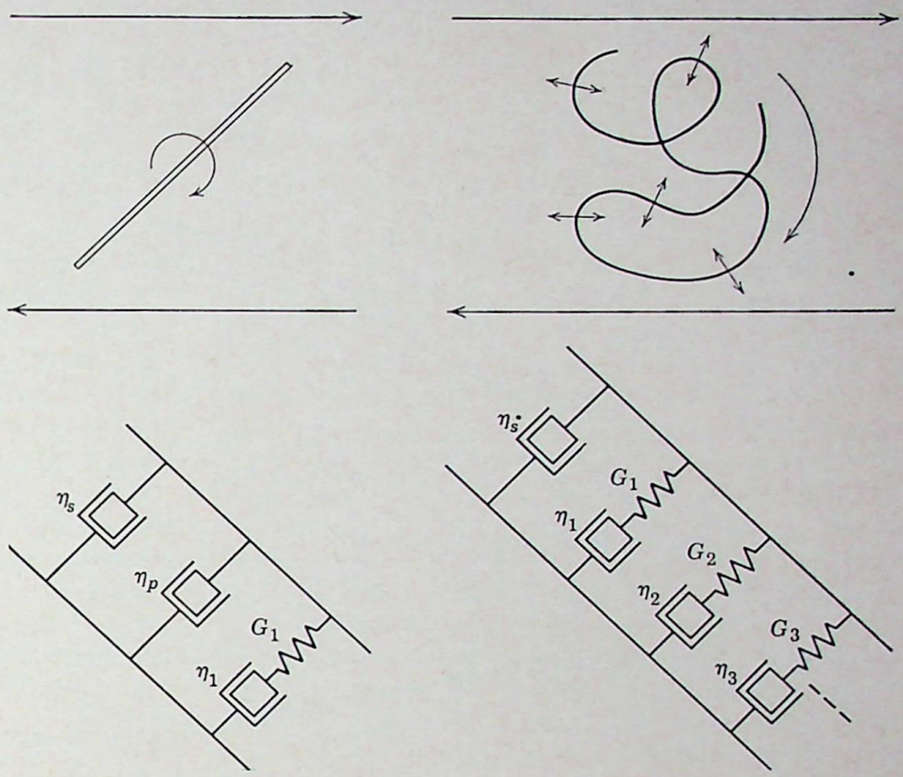


FIG. 1. Mechanical models corresponding to theoretical viscoelastic behavior of dilute rigid rods and flexible coils (below); schematic hydrodynamics (above).

lar weight, M, and the concentration, c (in g/cc). The values for rigid ellipsoids[6,7] are quite similar to those for rigid rods, and those for coils with internal hydrodynamic interaction[3] are somewhat similar to those for the free-draining type. Of course, the portrayal of a mechanical model is not an essential feature, but it aids visualization. The terminal relaxation time (for rods, the only relaxation time) is $\tau_1=\eta_1/G_1$.

Table I also gives the steady-state compliance for these systems. In very dilute solutions, it is proportional to the square of the intrinsic viscosity $[\eta]$ (expressed here in cc/g) and to the molecular weight. It is interesting that the values for rods and coils differ only by a minor numerical factor. In concentrated solutions (defined by the condition that $\eta \gg \eta_s$), J depends on the molecular weight only; if there are different macromolecular components, it is weighted strongly by those of highest molecular weight.[8]

Under the low stresses which would be necessary to avoid extreme non-Newtonian and non-Hookean effects, the magnitude of the stored elastic energy is quite small. For example, a dilute solution of tobacco mosaic virus[9] flowing under a shear stress of 10 dyne/cm^2 would store 25 cal/mole of solute; one of hyaluronic acid[10] (molecular weight 500 000), 0.25 cal/mole. At a concentration of 1%, the terminal relaxation times for these systems are calculated to be of the order of 10^{-3} and 10^{-4} sec, respectively, which are so short as to preclude direct experimental observation of a transient elasticity; macroscopic elastic recoil would be vitiated by the inertia of the solvent. The elasticity could be measured by dynamic experiments, however, as described in the following section. By contrast, the behavior of a concentrated solution of flexible coils may be illustrated by a 2% solution of the sodium salt of DNA, which[11] under a shear stress of 100 dyne/cm^2 would store about 200 cal/mole. (Sodium DNA molecules are of course not very flexible, but in concentrated solution exhibit some flexibility which may arise from hinges where the helix is unraveled for short distances.) The elastic shear would amount to about 50%, and the terminal relaxation time is calculated to be of the order of 1000 sec, so a substantial slow elastic recoil is readily observable following cessation of flow. It should be emphasized that to obtain such elastic effects, familiar in various biological systems (mucous and gelatinous secretions, slime molds), it is not necessary to have any linkages between the macromolecules or a netted gel structure.

Under higher stresses, substantial amounts of energy may be stored, but the complications of nonlinear effects prevent making numerical estimates. With a high degree of orientation, the structure of the system may be completely changed by the appearance of ordered phases,[4,5] and the changes may be irreversible as in the secretion of silk.[12] These phenomena may give some hint of the nature of rheological processes in protoplasm.[13]

III. ENERGY STORAGE AND DISSIPATION IN OSCILLATING MOTIONS

When a stress is applied to a macromolecular system, a finite time of the order of τ_1 is required for the stored energy to build up to its steady-state value. The changes during this transient period are determined by the characteristic relaxation times, or relaxation spectrum, mentioned in Sec. I. Equivalent information about the relaxation times can be obtained, often more easily, by subjecting the system to a sinusoidally oscillating shear

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

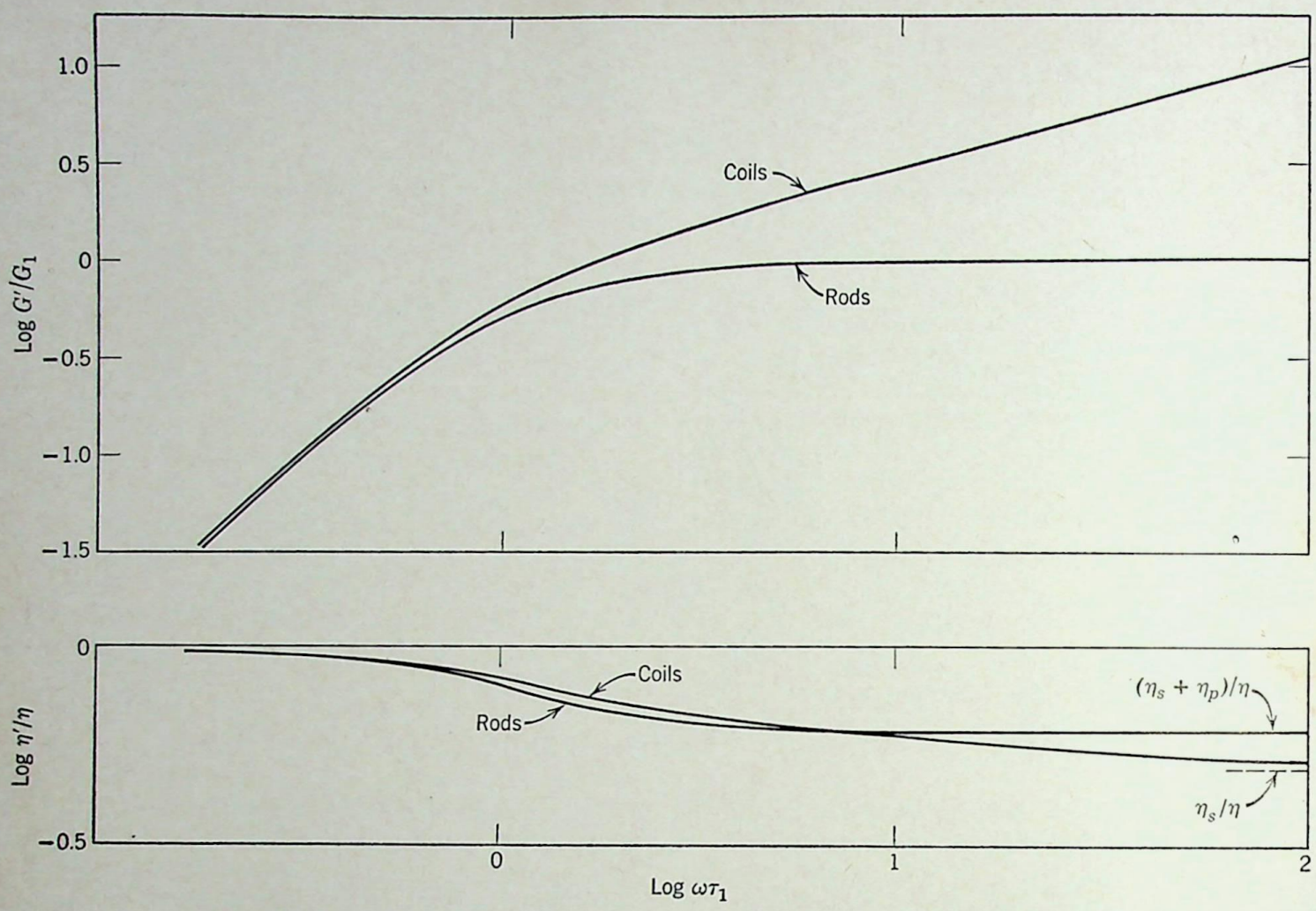


FIG. 2. Predicted frequency dependence of the dynamic storage modulus G' and the real part of the complex dynamic viscosity η', for dilute rigid rods and free-draining flexible coils, at concentrations such that $\eta = 2\eta_s$.

stress. Here some of the energy is stored and released each cycle, and some is dissipated. The respective values (for a given peak strain) are proportional to the storage modulus G' and the real part of the dynamic viscosity η', which are defined as the stress divided by those components of strain and rate of strain, respectively, in phase with the stress. Although it is doubtful whether any rhythmic biological processes can be approximated by this pattern, sinusoidal experiments are very useful for deducing the rates of molecular rearrangements.

At very low frequencies, η' approaches the ordinary steady-flow viscosity η, and G' approaches zero. However, the limiting value of $G'/(\omega\eta')^2$, where ω is the radian frequency, remains finite at vanishing ω and is in fact the steady-state compliance J. With increasing frequency, G' increases and η' decreases monotonically. These changes correspond to the obvious expectation in the mechanical models of Fig. 1 that with increasing frequency the motions will occur more in the springs and less in the dashpots. On a molecular scale, they mean that, as the period of oscillation is shortened, there is less opportunity for Brownian motion (rigid rotatory diffusion for rods, configurational rearrangements for coils) to erase the orientation imposed by the external deformations.

The characteristic frequency dependence expected for two simple cases is shown in Fig. 2, calculated for dilute solutions at a concentration such that $\eta = 2\eta_s$. For rigid rods,[1] there is a single relaxation time; G' should approach a limiting high-frequency value of G_1, and η' should approach $\eta_s + \eta_P$ (cf. Fig. 1 and Table I). Measurements of transverse wave propagation in solutions of rod-like intermediate polymers of fibrinogen are in rough agreement with this prediction.[14] For free-draining flexible coils,[2] there is a spectrum of relaxation times reflecting various degrees of cooperation between neighboring segments in configurational rearrangements; η' should decrease and G' increase more gradually with increasing frequency, following slopes of -0.5 and 0.5, respectively, on a doubly logarithmic plot, as the motions responding within the period of oscillation become restricted to shorter range cooperation. Eventually η' should approach η_s, and G' should attain a very high value of the order of the rigidity of a hard glass multiplied by the volume fraction of solute (this limit, which is not taken into account in the theories quoted,[2,3] corresponds to complete lack of configurational rearrangement within the period of oscillation, and is probably of no relevance to biological conditions). At present there are no data of this sort on polymers of biological origin; measurements on dilute solutions of synthetic polymers follow the general course shown in Fig. 2, although a somewhat different slope would be expected because of hydrodynamic interaction.[3]

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

Concentrated, rather than dilute, solutions are of primary interest in biological systems. Here the frequency dependence of dynamic mechanical behavior is more complicated, as shown in Fig. 3 for a 1.2% solution of sodium DNA.[11] Although the coils of this polymer have perhaps only moderate flexibility, the behavior follows a well-established pattern common to many flexibly coiling polymers of high molecular weight in concentrated solution.[15] There is an intermediate frequency range in which η' falls steeply and G' changes relatively little. This is associated with long-range coupling phenomena which cause the motions of one molecule to influence those of others separated by considerable distances but connected by some kind of chain of entanglements or coupling points.

The intermediate region—often called the plateau zone—lies between two extreme regions where the mechanical behavior is dominated by quite different physical processes. At high frequencies (corresponding to short times in transient experiments), the relaxation times are determined by the local friction of a short coil segment moving through its surroundings, oblivious of long-range coupling. The frequency dependence of G' and η' follows the Rouse theory for free-draining flexible coils, and from it the magnitude of the local friction can be calculated (Sec. IV). At low frequencies (corresponding to long transient times), the relaxation times are determined by the long-range coupling, which in turn is reflected in the steady-flow viscosity (Sec. V).

All the oscillatory phenomena described here relate to macroscopic deformation in shear. High-frequency measurements of G' and η' can be made by wave-propagation experiments provided shear waves are employed. In most sonic and ultrasonic wave-propagation experiments, however, the waves are longitudinal and the deformation is a combination of shear plus bulk compression. The energy storage is almost entirely due to the compression, in which, for a polymeric solution, the solute would play a minor role.

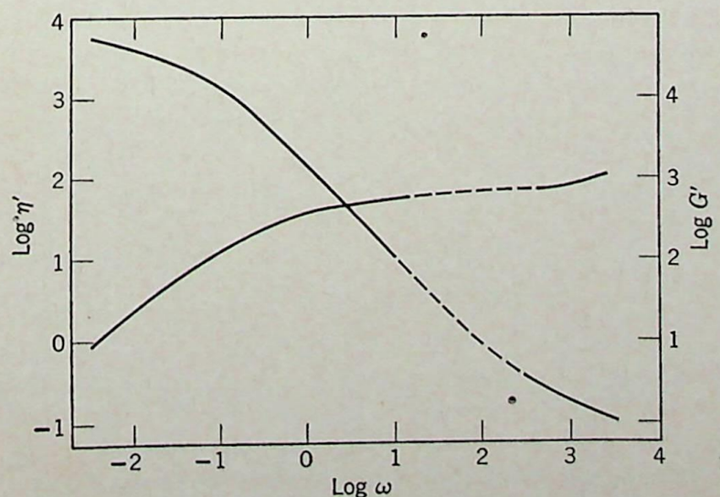


FIG. 3. Observed frequency dependence of G' (ascending curve) and η' (descending curve) in a 1.2% solution of sodium DNA.[11]

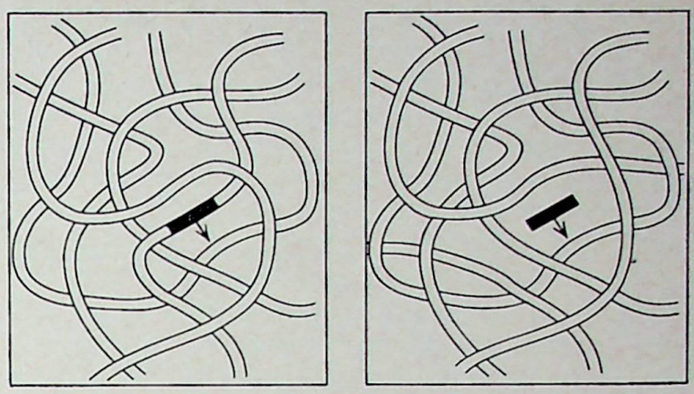


FIG. 4. Schematic representation of translational friction for a unit of a macromolecular chain (left) and a small foreign molecule in a macromolecular matrix (right).

IV. LOCAL MOLECULAR FRICTION

The frictional forces encountered by a short segment of a flexible coil as it pushes its way through its surroundings in a concentrated solution are very complicated but can, on the average, be described by a friction coefficient per monomer unit, ζ_0, which is force per unit velocity. This represents the same sort of translatory friction which a foreign molecule encounters in diffusion (Fig. 4). For certain soft synthetic polymers (with no diluent present), it has been possible to compare the friction coefficient per monomer unit of the wriggling chain, calculated from rheological measurements such as those in Fig. 3, with the friction coefficient of a foreign molecule similar in size to the monomer unit, obtained from diffusion measurements where only a trace of the foreign component is present.[16] They are actually of similar magnitude—for example, in polyisobutylene at 25°C, 4.5×10^{-5} dyne sec/cm for the chain unit, and 3.8×10^{-5} for a pentane molecule.

The friction coefficient reflects a sort of local viscosity which involves pushing aside solvent molecules together with short segments of other polymer coils in the immediate vicinity. A very crude estimate of the effective local viscosity η_e can be obtained from Stokes' law ($\zeta_0=6\pi\eta_e r$), taking r as the radius of a sphere with the volume of a monomer unit. When the solution is not highly concentrated, the limited information available indicates that η_e is not far from the solvent viscosity,[17] and is thus much smaller than the macroscopic steady-flow viscosity. With increasing polymer concentration, ζ_0 and η_e increase rapidly, but not nearly as rapidly as does the macroscopic viscosity η. The most extreme divergence is reached in an undiluted soft polymer; for the polyisobutylene quoted above, for example, η_e is about 70 poises and η is 10^{10}. For sufficiently high molecular weights, ζ_0 and η_e are independent of molecular weight, since the local motions are oblivious of the distant ends of the macromolecules. But η increases very rapidly with molecular weight, often with the 3.4 power.[18]

It is evident that the effective local viscosity encountered in translatory friction of a foreign molecule

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

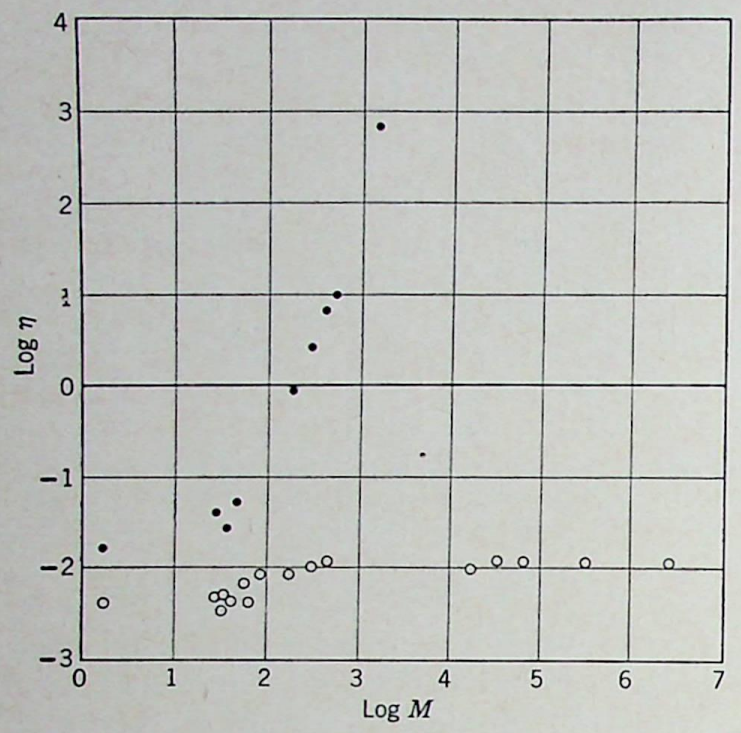


FIG. 5. Local effective viscosities of small organic molecules in soft rubber (black points) and in water (open circles), plotted against their molecular weights on a doubly logarithmic scale (after F. Grün[22]).

should increase continuously with molecular size as its motion involves increasingly long-range cooperation of the coil segments which rearrange around it. When the size reaches a magnitude such that long-range coupling affects its motion, there should be a large increase in η_e, and eventually for macroscopic dimensions of the moving particle η_e must approach η.

There are some fragmentary examples of a striking dependence of local effective viscosity on the size of a moving particle, derived from measurements of centrifugation and diffusion in matrices of flexible coils. Bushy stunt virus molecules sedimented in an approximately 0.1% solution of sodium DNA encountered an effective viscosity only 20% higher than that of the solvent, in studies by Schachman and Harrington,[19] while much larger polystyrene latex particles encountered a viscosity higher by a factor of 30. A size dependence for the effective viscosity of latex particles of different diameters sedimenting through concentrated sodium-polyacrylate solutions has been found by Ferry and Morton.[20] Diffusion of sucrose in concentrated aqueous solutions of polyvinyl pyrrolidone, studied by Nishijima and Oster,[21] showed that the local effective viscosity encountered by the sucrose was independent of the molecular weight of the polymer and far smaller than the macroscopic viscosity. Finally, measurements of diffusion of various organic compounds, with molecular weights ranging up to 1000, in undiluted soft rubber enabled Grün[22] and Kuhn[23] to calculate effective local viscosities which increased very rapidly with molecular size, as shown in Fig. 5. (The figure also shows effective viscosities calculated from diffusion measurements in water, which, by contrast, simply correspond to the macroscopic viscosity of water over the whole molecular-weight range.)

The magnitude so fall relaxation times reflecting configurational changes of the sort described by the Rouse theory are proportional to ζ_0. With increasing temperature, ζ_0 falls and the relaxation spectrum shifts to shorter times without changing its shape; curves such as those in Fig. 3 shift to higher frequencies without changing their shapes. For biological systems, temperature changes are probably much less important than concentration changes. With a decrease in polymer concentration, which might arise from osmotic flow, ζ_0 could drop sharply, speeding up all molecular rearrangements and conceivably allowing stored mechanical energy to be dissipated or to accomplish macroscopic movements.

V. LONG-RANGE COUPLING PHENOMENA

The plateau zone in Fig. 3, observed in all high molecular-weight polymers in concentrated solution or the undiluted state, represents qualitatively the behavior to be expected if the macromolecules were coupled together quite strongly, though slipping to some extent, at widely separated points.[24] The same concept of long-range coupling explains another widely observed phenomenon in the dependence of the steady-flow viscosity on molecular weight.[25] A doubly logarithmic plot of viscosity against molecular weight or degree of polymerization shows a sharply increased slope above a critical value of the abscissa, which in typical cases corresponds to $M = 10\,000$ to $50\,000$ (Fig. 6). The long-range coupling cannot, in general, be attributed to secondary bonding at loci of attraction, because it appears in nonpolar polymers which are quite homogeneous in chemical composition. It is apparently re-

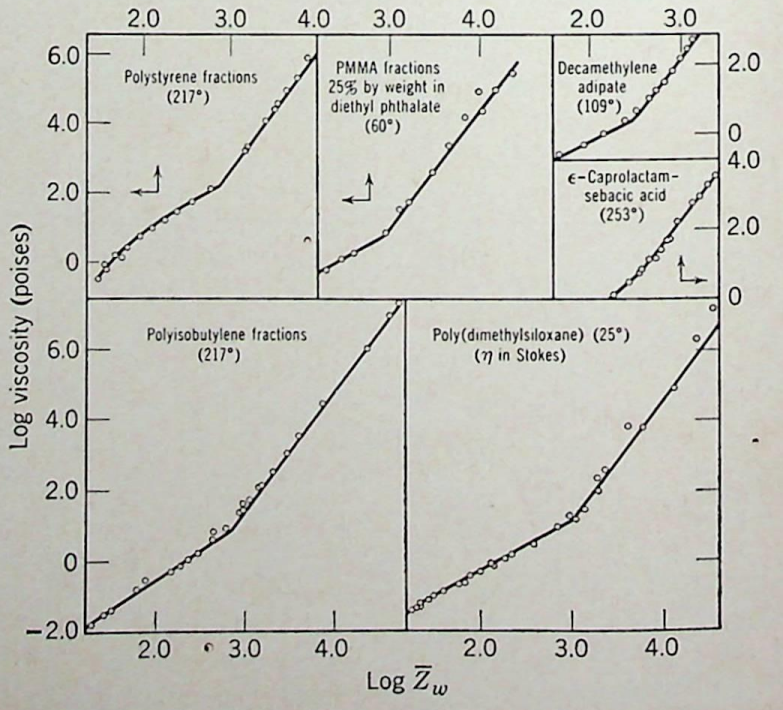


FIG. 6. Doubly logarithmic plots of viscosity against degree of polymerization for several types of synthetic polymers, showing proportionality to the 3.4 power of degree of polymerization or molecular weight above a critical value (after Fox and Loshaek[18]).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

lated to the expectation that a molecule, in its long-range contour, will form a complete loop enclosing similar loops of other molecules,[26] but the topological requirements are not clear. Its influence increases with increasing molecular stiffness, as evidenced by comparing synthetic vinyl polymers, cellulose derivatives, and DNA.[11]

The terminal relaxation time in a concentrated solution is determined by the long-range coupling, and the extent to which motions of one molecule can influence others, through chains of mutual entanglement. Since the steady-flow viscosity is also dominated by the coupling, however, the order of magnitude of τ_1 can still be estimated by the product $J\eta$.

Quite different from the above topological kind of coupling are the stronger linkages which cause some concentrated polymer solutions to gel. Here bonds of considerable permanence are formed, and the resulting network has no finite viscosity. It has no equilibrium elasticity either; under constant stress it continues to deform indefinitely at a very small and steadily decreasing rate. The spectrum of relaxation times extends indefinitely to long times, probably reflecting a mutual influence between molecular motions which extends throughout the system, communicated through the linkages. This is manifested, for low-frequency oscillating motions, by a very slight decrease in G' with decreasing frequency together with an increase in η' apparently without limit, as illustrated in Fig. 7 for a gel of polyvinyl chloride.[17] In some gels, the linkages are probably very small crystalline regions, tying together handfuls of flexible strands,[27] growth of the crystallites being limited by some kind of structural heterogeneity along the macromolecular chain. In others, there may be local secondary bonding of specific chemical groups.[28] As an extreme case, primary chemical bonds may join the macromolecules into a network, as proposed for gels of denatured serum albumin[29,30] and fibrin clots formed with fibrin-stabilizing factor.[31,32] Thus, while a netted structure is not essential to the storage of elastic energy under stress, it plays an important role when it is present.

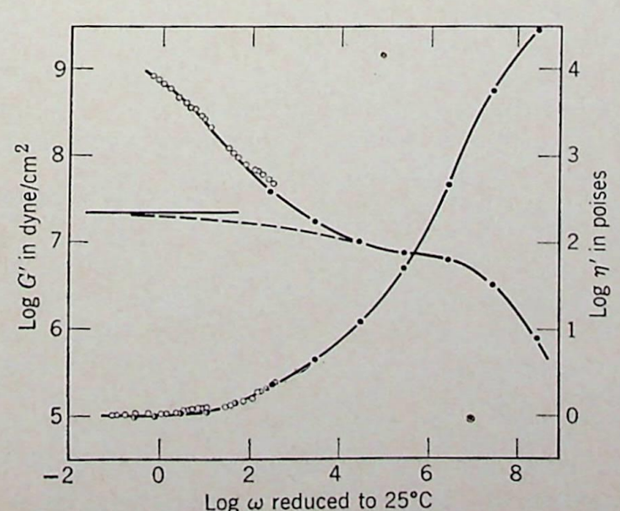


FIG. 7. Frequency dependence of G' and η' for a 10% gel of polyvinyl chloride in dimethyl thianthrene.[17]

The nature of molecular responses to very slow external motions in such networks, involving cooperation among widely separated network strands, is poorly understood. But if the strands between the linkage points are flexible macromolecular chains, the responses to rapid motions will be oblivious of the linkages and their rates will be governed by the local friction coefficient just as in the examples discussed in Sec. IV.

Changes in the consistency of protoplasm[13] suggest that within it network linkages are sometimes formed and sometimes broken. Such structural modifications should not affect responses to rapid motions but would enormously alter the relative storage and dissipation of energy in slow motions. Correspondingly, they should markedly influence frictional resistance to translation of large particles while leaving the resistance to small molecules essentially unchanged.

BIBLIOGRAPHY

[1] J. G. Kirkwood and P. L. Auer, J. Chem. Phys. **19**, 281 (1951).
[2] P. E. Rouse, J. Chem. Phys. **21**, 1272 (1953).
[3] B. H. Zimm, J. Chem. Phys. **24**, 269 (1956).
[4] G. Oster, J. Gen. Physiol. **33**, 445 (1950).
[5] P. J. Flory, Proc. Roy. Soc. (London) **A234**, 73 (1956).
[6] R. Cerf, Compt. rend. **234**, 1549 (1952).
[7] H. A. Scheraga, J. Chem. Phys. **23**, 1526 (1955).
[8] J. D. Ferry, M. L. Williams, and D. M. Stern, J. Phys. Chem. **58**, 987 (1954).
[9] M. A. Lauffer, J. Am. Chem. Soc. **66**, 1188 (1944).
[10] T. C. Laurent, J. Biol. Chem. **216**, 263 (1955).
[11] F. E. Helders, J. D. Ferry, H. Markovitz, and L. J. Zapas, J. Phys. Chem. **60**, 1575 (1956).
[12] K. H. Meyer and J. Jeannerat, Helv. Chim. Acta **22**, 22 (1939).
[13] L. V. Heilbrunn, *The Dynamics of Living Protoplasm* (Academic Press, Inc., New York, 1956).
[14] J. D. Ferry and F. E. Helders, Biochim. et Biophys. Acta **23**, 569 (1957).
[15] J. D. Ferry in *Die Physik der Hochpolymeren*, H. A. Stuart, editor (Springer-Verlag, Berlin, 1956), Vol. IV, Chap. VI.
[16] J. D. Ferry and R. F. Landel, Kolloid-Z. **148**, 1 (1956).
[17] J. D. Ferry, D. J. Plazek, and G. E. Heckler, J. chim. phys. **55**, 152 (1958).
[18] T. G. Fox and S. Loshaek, J. Appl. Phys. **26**, 1080 (1955).
[19] H. K. Schachman and W. F. Harrington, J. Am. Chem. Soc. **74**, 3965 (1952).
[20] J. D. Ferry and S. D. Morton, unpublished experiments.
[21] Y. Nishijima and G. Oster, J. Polymer Sci. **19**, 337 (1956).
[22] F. Grün, Experientia **3**, 490 (1947).
[23] W. Kuhn, Makromol. Chem. **6**, 224 (1951).
[24] J. R. McLoughlin and A. V. Tobolsky, J. Colloid Sci. **7**, 555 (1952).
[25] F. Bueche, J. Chem. Phys. **20**, 1959 (1952).
[26] F. Bueche, J. Polymer Sci. **25**, 243 (1957).
[27] J. T. Alfrey, Jr., N. Wiederhorn, R. Stein, and A. V. Tobolsky, J. Colloid Sci. **4**, 211 (1949).
[28] G. Stainsby, editor, *Recent Advances in Gelatin and Glue Research* (Pergamon Press, London, 1958).
[29] C. Huggins, D. R. Tapley, and E. V. Jensen, Nature **167**, 592 (1951).
[30] V. D. Hospelhorn, B. Cross, and E. V. Jensen, J. Am. Chem. Soc. **76**, 2827 (1954).
[31] A. Loewy and J. T. Edsall, J. Biol. Chem. **211**, 829 (1954).
[32] L. Lorand and A. Jacobsen, J. Biol. Chem. **230**, 421 (1958).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

16
Respiratory-Energy Transformation*

ALBERT L. LEHNINGER

Department of Physiological Chemistry, The Johns Hopkins School of Medicine, Baltimore 5, Maryland

INTRODUCTION

THIS article summarizes the broad outlines of current knowledge on some enzymatic and biophysical aspects of the conversion of respiratory energy into phosphate-bond energy. There are many problems for the biophysicist in this area which not only are pertinent to electron transport and respiratory-chain phosphorylation but also have wide applicability to more-general problems of ultrastructure and arrangement of enzymes in biologically organized "solid-state" arrays. More-detailed treatments of the subject can be found in review articles.[1–5]

ENERGY CYCLE IN AEROBIC CELLS

In all cells, the energy-requiring or *endergonic* functions are driven by chemical energy liberated during the energy-yielding or *exergonic* degradation of foodstuff molecules. The major energy-requiring functions or activities of cells include (a) *biosynthesis* of complex molecules from simple ones when such reactions proceed in an environment thermodynamically unfavorable for synthesis, as in the formation of glycogen from glucose in an aqueous system, (b) *mechanical work*, as in muscular contraction, (c) *active transport* or accumulation of substances against gradients of chemical potential, (d) *transmission* or conduction phenomena, and (e) *bioluminescence*. In aerobic cells, oxidative degradation of foodstuffs is the source of chemical energy. In strictly anaerobic cells, molecular oxygen does not intervene but internal oxido-reductions of fermentative or anaerobic metabolic cycles (such as alcoholic fermentation of glucose) are the primary energy source to drive endergonic functions.

The energy cycle in nature (as well as the carbon cycle) is then completed by the conversion of radiant energy of sunlight into chemical energy during photosynthesis, making possible reduction of carbon dioxide back to the oxidation level of carbohydrate, in reactions discussed in the companion paper by Calvin (p. 147). Respiratory-energy transformation and photosynthesis thus constitute the two complementary mainsprings of biological energy.

CONCEPT OF PHOSPHATE-BOND ENERGY

The general medium of energy exchange between certain exergonic reactions of oxidative catabolism and the endergonic functions is the so-called "high-energy phosphate bond" of ATP.

Phosphate esters of cells can be classed into two major groups on the basis of the magnitude of the standard free energy of hydrolysis of the phosphate bond.[6–8] Simple phosphate esters, such as 3-phosphoglyceric acid, glucose-6-phosphoric acid, and α-glycerophosphoric acid, etc., undergo hydrolysis enzymatically or by acid or base catalysis, often at widely different rates, with free energies of hydrolysis (ΔF^c) of -0.8–-3.0 kcal/mole at 25° and *p*H 7. These are the so-called "low-energy phosphate compounds." On the other hand, enol phosphate esters (phosphopyruvic acid), guanidine phosphates (phosphocreatine), acyl phosphates (1,3-diphosphoglyceric acid), and pyrophosphates (ATP) are in the high-energy class and have a $\Delta F°$ for hydrolysis of from 6 to 15 kcal/mole under standard conditions. This division is arbitrary and exact values of these esters are now in a state of flux because $\Delta F°$ values for hydrolysis of ATP are currently undergoing substantial revision.[9]

The *free energy of hydrolysis* represents the difference in energy content of reactants and hydrolysis products under standard conditions. It is *not* equivalent, of course, to the true bond energy or the zero-point energy; actually, input of energy is required to break chemical bonds. It must be made clear also that there is no particular biological magic involved in the difference between high- and low-energy compounds. Actually, the major reason for the relatively high free energy of hydrolysis of ATP and other high-energy compounds is that the immediate hydrolysis products of these compounds undergo secondary, spontaneous reactions with high equilibrium constants, leading to thermodynamically more-stable forms, either through resonance stabilization or tautomeric rearrangements. A second contribution to the high free energy of hydrolysis, in the case of ATP and phosphopyruvate, for example, is the separation and randomization of the neighboring like charges in these molecules at physiological *p*H. The structural, resonance, and charge contributions to the high-energy nature of these compounds have been discussed by Hill and Morales.[10] Classically, the free energy of hydrolysis of ATP has been assumed to have the value 12 kcal/mole, on the basis of a calorimetrically observed ΔH of approximately this figure and the assumption of only a small entropy change. However, recent work on the enzymatic equilibria of ATP and glutamate and the highly refined heat measurements by the important new method of Benzinger[11] have revealed that the standard free-energy change at

* Original experiments from the author's laboratory reported in this article were carried out with the aid of grants from the Public Health Service, U. S. Department of Health, Education, and Welfare, the National Science Foundation, The Nutrition Foundation, Inc., The Whitehall Foundation, and the Albert and Mary Lasker Foundation.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

1*M* reactants corrected for ionization changes at *p*H 7.5 is about −8.4 kcal, and that the true ΔH for hydrolysis is about −4.8 kcal (see Burton[9] for most recent evaluation). The value of the free energy of the hydrolysis of the ATP bond (−8.4 kcal) now seems to be a valid one; it is thus substantially lower than the classical value of 12 kcal. It must be pointed out, however, that, *under physiological conditions of pH, and intracellular concentration ranges of phosphate, ADP and ATP*, the recalculated $\Delta F'$ of ATP is about −12.5 kcal/mole or approximately the classical textbook value.[9] There is room for a great deal of careful work on the measurement of biochemically useful thermodynamic factors through refined heat measurements, enzymatic equilibria, and calculation of free energies of formation.[12,13] The values given above for ATP may be altered further when the hydrolysis is considered in a physiological milieu because of the lack of information on activity coefficients of the various ionic species of the reaction components—in particular, the Mg complexes of ATP and ADP in the presence of the high ionic strength background of intracellular fluid. Also, it must be pointed out that free-energy data for ATP obviously are reckoned on a macroscopic, statistical basis. In the intact cell, ATP-linked reactions occur with reactants bound to proteins (as in actomyosin) under nonequilibrium conditions where energy changes and transfers at the *microscopic* level dictate direction and rate of reactions.

There may be good reasons for the evolutionary choice of phosphoric acid, rather than sulfuric acid, HCl, or a carboxylic acid, as the vehicle for biological energy transfer through enzymatic group-transfer reactions. In the first place, phosphate anion has itself a rather high resonance energy; however, this property is not unique to phosphate. Perhaps it is more important that its anhydrides and amides are kinetically (as opposed to thermodynamically) much more stable than esters or amides of HCl, H_2SO_4, or carboxylic acids. Thus, acetic anhydride in H_2O at *p*H 7.0 decomposes rapidly, whereas the anhydride pyrophosphoric acid at the same *p*H has a much greater half-life despite the fact that the free energy of hydrolysis of the two compounds is of the same order of magnitude. Phosphate anhydrides and amides may have been selected biologically not only for their resonance characteristics but also for their property of remaining relatively unreactive in an aqueous solution in the absence of suitable directing enzymes; less-stable derivatives would tend to react spontaneously and thus not be subject to enzymatic control.

COMMON INTERMEDIATE PRINCIPLE IN UTILIZATION OF ATP ENERGY

Apart from certain photochemical and fluorescence phenomena, there is, in general, only one way in which energy can be transferred from one chemical reaction to another. The two reactions must be consecutive and *must have a common intermediate.*

The synthesis of sucrose provides a classical biochemical example.[14] The reaction

$$\text{glucose}+\text{fructose} \rightleftarrows \text{sucrose}+H_2O \quad (\Delta F=+5000) \qquad (1)$$

is strongly *endergonic*. Sugar-cane juice is 0.5*M* in sucrose. The concentrations of free glucose and fructose would have to be enormously greater than that of sucrose, and the water concentration low, in order for the reaction to proceed spontaneously in the sugar cane. This is not the case and sucrose is not made in this way. Enzyme studies show that it is formed at the expense of the hydrolysis of ATP in a coupled reaction. The driving reaction is the strongly *exergonic* hydrolysis of ATP:

$$\text{ATP}+H_2O \rightarrow \text{ADP}+\text{P} \quad (\Delta F=-12\,000). \qquad (2)$$

Reactions (1) and (2) are the partial thermodynamic reactions in sucrose synthesis. The actual enzymatic mechanism is the following:

$$\text{ATP}+\text{glucose} \longrightarrow \text{glucose}-1-P+\text{ADP} \quad (\Delta F=-7000 \text{ cal}) \qquad (3)$$

$$\text{glucose}-1-\text{P}+\text{fructose} \rightarrow \text{sucrose}+H_3PO_4 \quad (\Delta F=\sim 0 \text{ cal}). \qquad (4)$$

The over-all ΔF is −7000 cal and synthesis of sucrose proceeds spontaneously in the presence of the enzymes. In this sequence, glucose has been raised to the energy level of a glycoside by phosphorylation; glucose−1−P is now the common intermediate of two consecutive reactions which together have a negative net free-energy change, but in which a product of the first, *exergonic*, reaction has sufficient "group potential" in the form of the glycosyl phosphate to drive the second, *endergonic*, reaction to form sucrose.

In principle, this pattern underlies all endergonic biosynthetic reactions, such as formation of glycogen from glucose, proteins from amino acids, and complex lipids from fatty acids. ATP and an intervening phosphate-group transfer reaction represent the mode of energy transfer.

COMMON INTERMEDIATE PRINCIPLE IN BIOLOGICAL CONVERSION OF OXIDATIVE ENERGY INTO PHOSPHATE-BOND ENERGY

From the foregoing, the breakdown of ATP to ADP and P is necessary to drive endergonic functions. The ATP then is regenerated continuously from ADP and P at the expense of catabolic reactions, primarily those of oxidation-reduction reactions, in order to complete the energy cycle.

The regeneration of ATP from ADP and P is very largely a matter of the transfer of the energy liberated in certain highly exergonic oxidations of metabolism to a reaction resynthesizing ATP from ADP and P. Such

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

respiratory energy conversion also occurs by the principle of the common intermediate.

The best-understood example for this discussion is the enzymatic oxidation of glyceraldehyde−3−phosphate to 3−phosphoglyceric acid by DPN which occurs in the anaerobic breakdown of glucose to lactate, to which is coupled the synthesis of ATP from ADP and phosphate. The over-all reaction first can be broken down into two partial thermodynamic reactions. The first, the oxidation of an aldehyde to a carboxylic acid, is highly exergonic, and the second, formation of ATP, highly endergonic.

$$\mathrm{R{-}\underset{\underset{\mathrm{O}}{\|}}{CH}{+}DPN_{ox} \rightleftarrows R{-}\underset{\underset{\mathrm{O}}{\|}}{C}{-}O^{-}{+}DPN_{red}{+}H^{+}}$$

$$(\Delta F{+}{-}16\ \text{kcal}) \quad (5)$$

$$\mathrm{P{+}ADP \rightarrow ATP{+}H_2O}\ (\Delta F{=}{+}12\ \text{kcal}). \quad (6)$$

The energy liberated in the oxidation of an aldehyde to a carboxylic acid is recovered as ATP, with energy left to spare, the net driving force being some 4 kcal, as is shown in the following over-all reaction:

$$\text{glyceraldehyde-3-phosphate}{+}P_i{+}DPN_{ox} \rightleftarrows \text{1,3-diphosphoglycerate}{+}DPN_{red}{+}H^{+}, \quad (7)$$

$$\text{1,3-diphosphoglycerate}{+}ADP \rightleftarrows \text{3-phosphoglycerate}{+}ATP. \quad (8)$$

It is seen that the oxido-reduction of reaction (7) results in formation of 1,3-diphosphoglycerate, a high-energy mixed anhydride of a carboxylic acid and phosphoric acid. This is the intermediate which is common in the two reactions and acts as a high-energy phosphate donor in the formation of ATP [reaction (8)]. Although this reaction is well understood, it accounts for only a very small fraction of ATP synthesis in aerobic cells, but it is an extremely important model of how the great bulk of ATP may be formed during respiration.

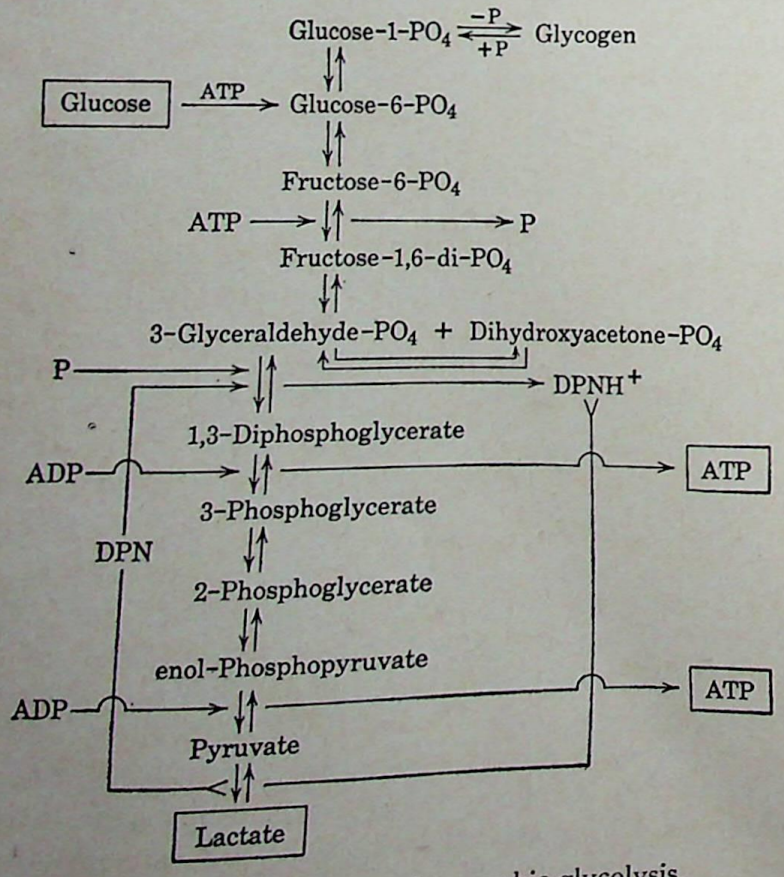


FIG. 1. The mechanism of anaerobic glycolysis.

RELATIVE CONTRIBUTION OF FERMENTATIVE AND OXIDATIVE REACTIONS TO ATP SYNTHESIS

Figures 1 and 2 show the outlines of the mechanism of anaerobic glycolysis and the Krebs citric-acid cycle, the main aerobic phase of carbohydrate oxidation. However, the amount of energy liberated from a glucose molecule during glycolysis to lactate ($\Delta F = -49\,700$ cal) represents a few percent only of the total energy liberated by complete oxidation of glucose to carbon dioxide and water ($\Delta F = -686\,000$ cal/mole). However, this small yield of energy in the anaerobic phase is quite efficiently recovered by the two phosphorylations occurring in the cycle, and from the following equations it is seen that some 50% of the energy liberated when 2 moles of lactate are formed from 1 of glucose is recovered in the form of the phosphate-bond energy of two moles of ATP.

Over-all reaction:

$$\text{glucose}{+}2ADP{+}2P \rightarrow 2\ \text{lactate}{+}2ATP \quad (\Delta F = -25\,700\ \text{cal}).$$

Partial reactions:

(a) glucose → 2 lactate ($\Delta F = -49\,700$ cal)

(b) $2ADP{+}2P \rightarrow 2ATP$ ($\Delta F = +24\,000$ cal).

However, the real energetic mainspring of aerobic cells, regenerating 90% or more of the ATP required, is the oxidative phase of catabolism.

Despite the absence of phosphorylated intermediates in the Krebs cycle as it is depicted in Fig. 2, it is known now that a very large number of phosphorylations of ADP accompany such aerobic oxidations, approximately 36 per mole of glucose oxidized. In addition, the efficiency of energy recovery during this aerobic phase is even higher than in the anaerobic phase, approaching some 70%.

OXIDATIVE PHOSPHORYLATION (SYNONYMS: AEROBIC PHOSPHORYLATION, RESPIRATORY-CHAIN PHOSPHORYLATION)

This term denotes the coupled phosphorylations of ADP which occur simultaneously with certain organized oxido-reduction reactions in the aerobic phase of metabolism. The biological significance of phosphorylations causally coupled to oxidations with molecular oxygen first was recognized by Kalckar in 1937, but the quantitative importance of these in cellular metabolism was not appreciated fully until the penetrating studies of Belitser in Russia in 1939–1941, which were

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

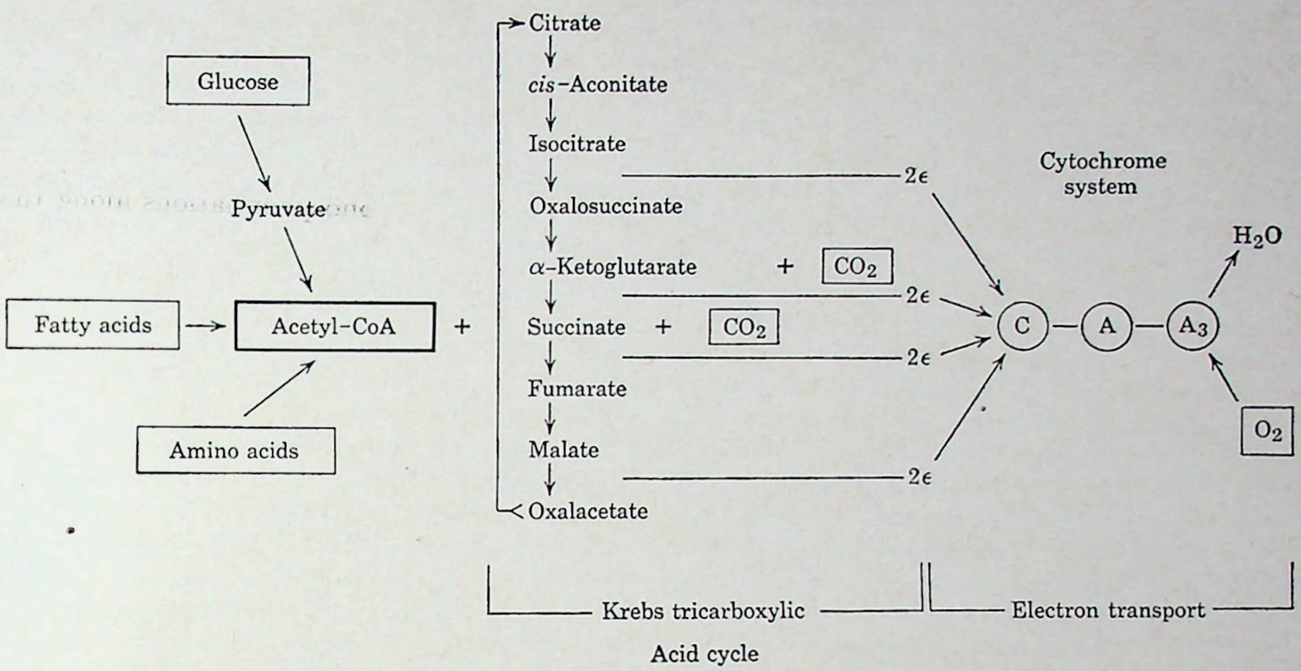


FIG. 2. The Krebs tricarboxylic-acid cycle.

followed during and after World War II by a number of independent and confirmatory studies by Ochoa, Cori, Hunter, and others (cf. references 4, 2, 3). These studies clearly demonstrated that, for the passage of every pair of electrons from a substrate of the Krebs citric-acid cycle to molecular oxygen during aerobic metabolism, an average of three moles of ATP were synthesized from ADP and P_i; i.e., the P:O ratio (moles P_1 taken up per atom oxygen used) was 3.0. Thus, the over-all equation for the oxidation of pyruvate to carbon dioxide and water as it occurs in the Krebs cycle was found to be

$$CH_3COCOOH+15P+15ADP+5O \rightarrow 3CO_2+15ATP+2H_2O. \quad (9)$$

This reaction may be broken down as follows:

$$CH_3COCOOH+5O \rightarrow 3CO_2+2H_2O \quad (\Delta F=-273 \text{ kcal}), \quad (10)$$

$$15P+15ADP \rightarrow 15ATP \quad (\Delta F=12\times15) \quad (\Delta F=+12\times15=180 \text{ kcal}), \quad (11)$$

$$\text{efficiency}=180\times100/273=67\%.$$

GENERAL PROPERTIES OF OXIDATIVE PHOSPHORYLATION

Oxidative phosphorylation classically has been recognized as an unstable and evanescent phenomenon which occurs only with relatively "native" and unfractionated extracts or homogenates of tissues. Aging of such extracts, or application of some of the common procedures for separation or purification of enzymes, quickly inactivated the coupled phosphorylation of ADP, whereas the purely oxidative activity of Krebs-cycle reactions remained rather stable. The findings thus suggested that coupled phosphorylation of ADP is not an obligatory accompaniment of oxidation, that it may be inactivated or uncoupled. The great lability of the coupling mechanisms to even relatively mild separation procedures has been a major technical obstacle to a fuller understanding of the mechanism of oxidative phosphorylation.

The great lability just described is explained in part by the fact that oxidative phosphorylation, along with the organized Krebs-cycle reactions and fatty-acid oxidation cycle, has been found to occur in rather highly organized subcellular structures, *the mitochondria*, as was first shown by Kennedy and Lehninger,[15] and independently by Schneider (cf. Ernster and Lindberg[16]), a finding which very quickly was confirmed and extended in many other laboratories.[17] Mitochondria isolated by the well-known sucrose procedure have become the standard study object in this field. It is important that mitochondria from many different cell types—plant, animal and microbial—show the respiratory-chain phosphorylation, thus providing a sound enzymatic basis for the suggestion by Claude that mitochondria are the power plants of the cell.

The mitochondria are almost completely self-contained units capable of carrying out both respiration and phosphorylation. It is necessary only to supplement a suspension of mitochondria with a substrate such as pyruvate, a trace amount of a 4-carbon catalyst such as fumarate or oxalacetate as substrates or fuel for the cycle, inorganic phosphate and ADP, and also Mg^{++}, which help to preserve mitochondrial structure. All of the necessary coenzymes and other co-factors are contained in the mitochondrial structure in such an organized manner that the complex cycle oxidations are smooth and complete, with no obvious or conspicuous bottle-

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

necks, indicating the high degree of enzymatic and morphological organization in the mitochondria.

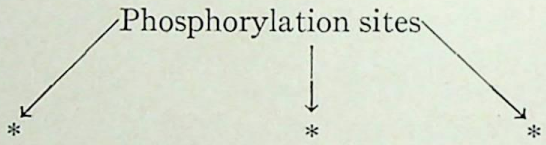
The studies on oxidative phosphorylation have revealed, however, that relatively intact mitochondrial structure is necessary; simple procedures, such as freezing and thawing, treatment with detergents, exposure to hypotonic or strongly hypertonic agents, or solution, all cause uncoupling of the phosphorylation.[4]

Another important property of oxidative phosphorylation is that certain inhibitors in relatively minute concentrations are capable of uncoupling phosphorylation; that is, in their presence, no phosphorylation of ADP takes place, but oxidation proceeds nearly normally. These uncoupling agents include 2,4-dinitrophenol, as well as other nitro- and halo-phenols, the antibiotic gramicidin, and the anticoagulant Dicumarol, in addition to a number of oxidation-reduction dyes such as methylene blue. These agents have become extremely important tools for the elucidation of the character and sequence of the intermediate reactions of oxidative phosphorylation.[18]

LOCALIZATION OF PHOSPHORYLATION SITES IN RESPIRATORY CHAIN[1-4]

To explain the occurrence of these oxidative phosphorylations, it was suggested at first that the intermediates of the Krebs cycle actually might occur in the form of phosphorylated derivatives rather than the free acids, analogs to the glycolytic reactions. Such phosphorylated intermediates never have been found. In 1940, however, Belitser postulated that the aerobic phosphorylations did not occur actually in the Krebs-cycle reactions proper, but rather during the phenomenon of electron transport from primary dehydrogenase to molecular oxygen via the known electron carriers making up the respiratory chain, such as diphosphopyridine nucleotide, flavoprotein, and the cytochromes. Belitser pointed out that the ΔF for the transfer of a pair of electrons from reduced diphosphopyridine nucleotide to molecular oxygen is approximately $-55\,000$ cal/mole. Since formation of one mole of ATP requires the input of 12 000 cal, it is evident that rather more than 4 moles of ATP theoretically could be generated during transport of each pair of electrons to oxygen, providing suitable enzymatic coupling mechanisms existed. As logical as the Belitser hypothesis appeared, experimental proof for this pattern was difficult to muster, and it was not until 1949–1951 that direct proof was adduced.[19,20] Highly purified, chemically reduced diphosphopyridine nucleotide was used as the actual substrate for the oxidation, in suspension of rat-liver mitochondria which had been pretreated in a hypotonic medium to produce sufficient alteration in permeability of the mitochondrial membrane so as to admit the reduced pyridine nucleotide to the oxidative centers within the mitochondria. Oxidation of reduced DPN via the known components of the respiratory chain, including the flavoprotein and cytochromes, culminated in oxygen uptake and coupled uptake of inorganic phosphate and the formation of ATP, with P:O ratios approaching the value 3.0. Oxidative phosphorylation thus occurs along the respiratory chain and may proceed in the absence of Krebs-cycle intermediates.

Similar experiments with pure cytochrome-*c*, used as either electron donor or acceptor, make possible further localization of the phosphorylations along the respiratory chain. Thus, passage of electrons from reduced DPN to cytochrome-*c* caused the formation of two moles of ATP.[4] Also, despite earlier experiments of Slater to the contrary, it finally was possible to demonstrate conclusively that the passage of a pair of electrons from cytochrome-*c* to oxygen yields a single phosphorylation.[21] More-recent spectroscopic investigations of Chance have confirmed these gross locations of phosphorylation sites.[1] Thermodynamic considerations of the known oxidation-reduction potentials of the carrier (Fig. 3), together with enzymatic and spectroscopic methods, suggest that one of the phosphorylations occurs between DPN and flavoprotein, the second between cytochrome-*b* and cytochrome-*c*. The site of the third phosphorylation is less certain: Chance suggests it lies between cytochrome-*c* and -*a*,[1] but other considerations indicate that it occurs between cytochrome-*a* and -a_3. These findings are summarized as follows:



$$\text{DPN} \rightarrow \text{flavoprotein} \rightarrow \text{cyt } b \rightarrow \text{cyt } c \rightarrow \text{cyt } a \rightarrow \text{cyt } a_3 \rightarrow \text{O}$$

The over-all reaction of the respiratory chain is thus

$$\text{DPN}_{red} + \text{O} + 3\text{P} + 3\text{ADP} \rightarrow \text{DPN}_{ox} + \text{H}_2\text{O} + 3\text{ATP}, \quad (12)$$

and may be considered as the partial reaction

$$\text{DPN}_{red} + \text{O} \rightarrow \text{DPN}_{ox} + \text{H}_2\text{O} \quad (\Delta F = -55\,000 \text{ cal}) \quad (13)$$

$$3\text{P} + 3\text{ADP} \rightarrow 3\text{ATP} + \text{H}_2\text{O} \quad (\Delta F = +36\,000 \text{ cal}). \quad (14)$$

From these results, it can be seen at once that the multimembered respiratory chain must represent a device for breaking up the large chunks of oxidative energy into smaller, biologically useful packets of approximately 12 kcal to accommodate the dimensions of the energetic currency of the cell—namely, the free energy of formation of ATP from ADP and phosphate. If all 55 kcal were released in one chemical-reaction step, even if coupled to a phosphorylation mechanism, only one ATP could be generated in this fashion.

MECHANISM OF RESPIRATORY-CHAIN PHOSPHORYLATION

The mechanism of oxidative phosphorylation, in enzymatic and chemical terms, remains as one of the

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

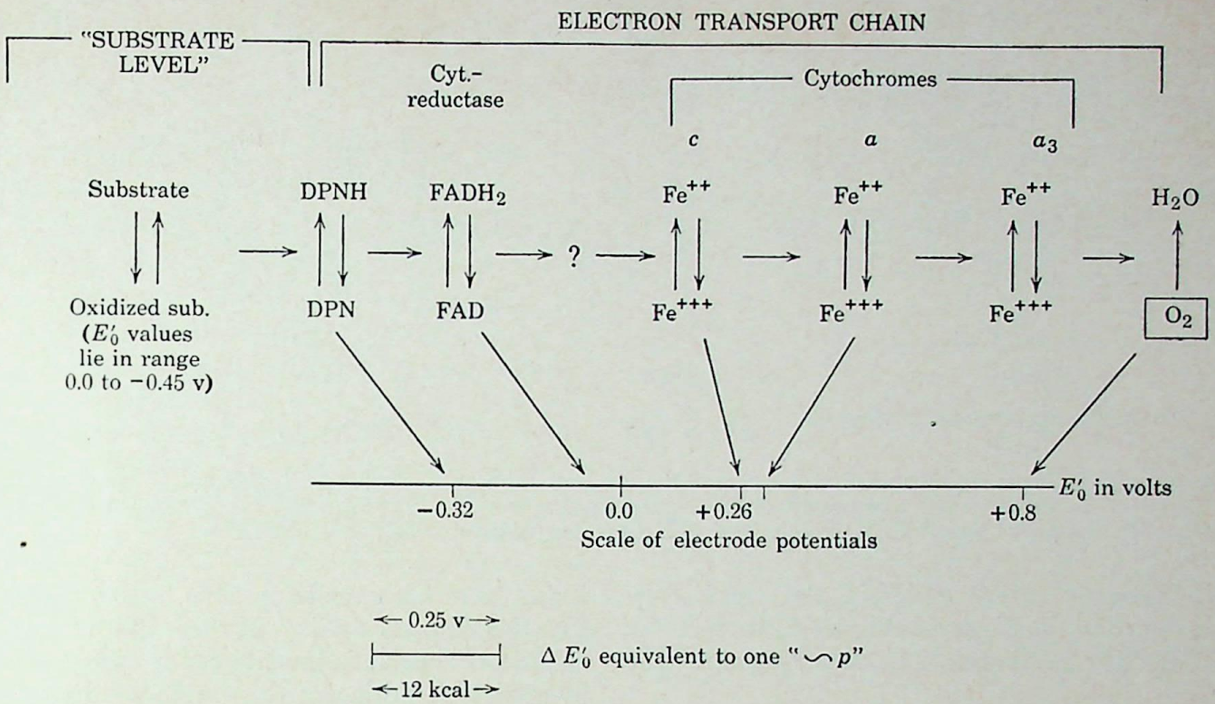


FIG. 3. The thermodynamic relationships in the respiratory chain. Electrons normally flow from substrates at the left to oxygen at the right. The approximate oxidation-reduction potentials (pH = 7.0) of the carriers are shown on the scale. The free energy of hydrolysis of ATP of about 12 kcal is equivalent to the maximum free-energy change as a pair of electrons flow over a 0.25-v span, as is shown in the yardstick, from the relation $F = nf\Delta E$ [from A. L. Lehninger in *Harvey Lectures* (Academic Press, Inc., New York, 1953–1954), Vol. IL, p. 176].

most conspicuous and challenging unsolved problems of contemporary biochemistry. The immediate situation is that the classical approach of separation and isolation of enzymes making up a complex metabolic sequence, followed by *in vitro* reconstruction of the enzyme system, simply has not been possible to date, not only owing to the fact that the process is very labile but also because there is a requirement for geometrical organization of the many enzymes concerned in the characteristic morphology of the mitochondria. In addition, there is the fundamental dilemma of how a set of respiratory carriers of even moderate substrate specificity can catalyze both nonphosphorylating and phosphorylating electron transport, particularly in view of the fact that highly purified respiratory carriers, such as cytochrome-*c*, DPN, cytochrome-*c* reductase, and other respiratory carriers, show absolutely no ability to phosphorylate when tested in highly purified form in the test tube [cf. reference 3].

To facilitate discussion of the problem of the mechanism of oxidative phosphorylation and some recent successes in unraveling part of the coupling mechanism, it is best to consider first the general mechanisms which have been proposed to account for energy coupling in the respiratory chain, which again make use of the principle of the common intermediate described below.

The following mechanism is one we have proposed and for which considerable evidence has accumulated.[3,22,5] It incorporates ideas originally proposed by Lipmann.[8] Similar mechanisms have been discussed since by Lardy, Hunter, Slater, Chance, and others (see references 2, 1, 23). In skeleton form, it is

$$\text{carrier}_{red} + X + \text{oxidant} \rightarrow \text{carrier}_{ox} \sim X + \text{reductant}, \quad (15)$$

$$\text{carrier}_{ox} \sim X + P_i \rightleftarrows \text{carrier} + P \sim X, \quad (16)$$

$$P \sim X + ADP \rightleftarrows ATP + X. \quad (17)$$

In this scheme, the substance X is the vehicle of energy coupling and forms a high-energy compound with the carrier during the course of transfer of electrons from the reduced carrier to the next in the chain (designated as "oxidant"). $\text{Carrier}_{ox} \sim X$ then undergoes phosphorolysis to form $P \sim X$ (a high-energy compound, presumably a phosphoenzyme) which in turn can donate its phosphate to ADP. Thus, carrier $\sim X$ is a common intermediate shared by the exergonic oxido-reduction of reaction (1) and by the endergonic reaction leading to formation of ATP in reactions (2) and (3). The energy liberated in the oxido-reduction thus is utilized to cause the formation of ATP. It is possible that phosphate interacts with the carrier directly to form a high-energy carrier~phosphate compound; however, for a number of reasons, it appears more likely that some compound other than phosphate is the first reactant with the carrier in the coupling mechanism. This hypothesis is written in skeleton form to indicate in the simplest way the principle of the common intermediate. It is possible and likely, however, that there are additional intermediate reactions between reactions (1) and (2) and also between (2) and (3).

This sequence now may be visualized as occurring

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

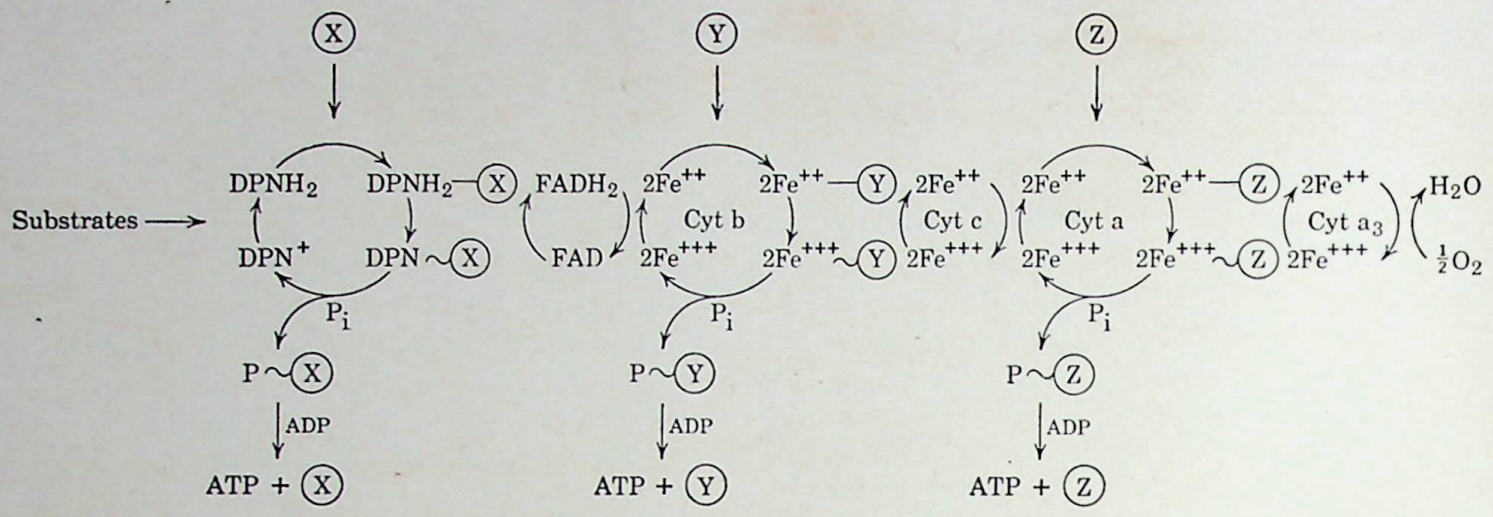


FIG. 4. A diagrammatic conception of the phosphorylation-coupled respiratory chain. This diagram is based on the coupling mechanism described in the text and assumes that the coupled carriers are DPN, cytochrome-*b*, and cytochrome-*a*. [from A. L. Lehninger, C. L. Wadkins, C. Cooper, T. M. Devlin, and J. L. Gamble, Jr., Science 128, 450 (1958)].

with three different carrier pairs in the respiratory chain, as is shown in the diagram (Fig. 4) to account for the three different phosphorylations occurring along the chain.

The postulated skeleton-reaction mechanism indicates two possible approaches to unraveling the mechanism of energy coupling in oxidative phosphorylation. The first and seemingly more direct approach is to carry out investigations to isolate and to identify the molecular species of the respiratory carrier which exists in the energy-charged form—in short, the structure and enzymatic reactions of carrier ~X or its analogs. However, not all of the known respiratory carriers have been obtained so far in purified soluble form suitable for examination of possible complexes of this kind and those which have been isolated to date have shown no propensity to phosphorylate in the test tube in simple systems. Furthermore, there now is increasing evidence that additional factors, such as vitamin K_1, α-tocopherol, quinones, copper, nonheme iron, and other substances capable of undergoing reversible oxidation-reduction changes, may be members of the respiratory chain in addition to the classically accepted flavoprotein and cytochromes, but the site of these in the chain is still quite uncertain. Direct identification of the coupled carriers as a first step to determining the identity of X also is made difficult by the probability that a coupled form such as carrier ~X reasonably could be expected to be rather labile. However, the spectroscopic studies of the kinetics of interaction of the carriers in the chain carried out by Chance in intact mitochondria provide some important landmarks for such efforts.[1]

There is another approach to the mechanism of oxidative phosphorylation, which sometimes is called the "back-door approach," and which we have chosen to take in our recent investigations.[5,18,24] The back-door approach starts with the end product of oxidative phosphorylation, namely ATP, and works back through the partial reactions to the carrier level. Our recent approaches to the mechanism of oxidative phosphorylation have been made possible by two developments. The first was our finding, in 1955, that phosphorylating submitochondrial fragments can be prepared from rat-liver mitochondria by the action of digitonin. These fragments are very much smaller than mitochondria but still retain phosphorylative activity. They are not simply miniature mitochondria since they no longer are capable of the Krebs-cycle reactions or of fatty-acid oxidation. However, these fragments, which are believed to be derived from the mitochondrial membrane, contain more or less intact respiratory chains together with the enzymatic factors responsible for energy coupling. Study of coupling in these so-called "digitonin fragments" is not complicated by the morphological compartmentation and permeability effects seen in intact mitochondria, which also catalyze many enzymatic reactions which are extraneous to oxidative phosphorylation.[25–31,22,5]

The second development is the finding that these particles catalyze two different isotopic exchange reactions of ATP which occur in the absence of electron transfers. The first is the ATP-P_i^{32} exchange, in which labeled P_i^{32} is rapidly incorporated into the terminal phosphate group of ATP.[22] The second is the so-called ATP-ADP exchange, in which P^{32}- or C^{14}-labeled ADP is incorporated bodily into ATP.[18] Both of the reactions in fresh fragments are completely inhibited by the classical uncoupling agent DNP, indicating their relationship to oxidative phosphorylation and distinguishing them from many other similar exchange reactions of ATP which have nothing to do with oxidative phosphorylation.

Close study of the interrelationship of these exchange reactions has revealed that the ATP-ADP exchange reaction is in reality the terminal reaction of oxidative phosphorylation

$$P\sim X + ADP \rightleftarrows ATP + X,$$

and is itself an intermediate step in the ATP-P_i^{32} exchange. Since phosphate is not involved in the ADP

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

exchange, but ADP is an obligatory component of the P_i^{32} exchange, the sequence of the two terminal reactions of oxidative phosphorylation is established as

$$R\sim X+P_i \rightleftarrows P\sim X+R$$
$$P\sim X+ADP \rightleftarrows ATP+X$$

where R may be an electron-carrier molecule. Further, it has been found that the terminal reaction is not itself inherently sensitive to DNP but that sensitivity to DNP is conferred on it because it is in equilibrium with the preceding reaction, which is DNP-sensitive.[5,18]

These experiments have led very recently to the separation of the enzyme catalyzing the ATP-ADP exchange in soluble form from the mitochondrial fragments. It has been purified by chromatography on cellulose derivatives.[5] The nature of the active site participating in formation of the phospho-enzyme intermediate is now under investigation.

With this terminal enzyme in hand as a foundation for reconstruction approaches, we now propose to work nearer the carrier level and hope to soon obtain definitive information on the oxidation-reduction state of the energy-rich form of the carriers[1,5,32] which drives the phosphorylation, as well as the identification of the three coupled carriers in the chain.

RESPIRATORY-ENZYME ASSEMBLIES AND THE STRUCTURE OF THE MITOCHONDRION

Consider now oxidative phosphorylation in the context of its natural habitat—namely, the mitochondrion.

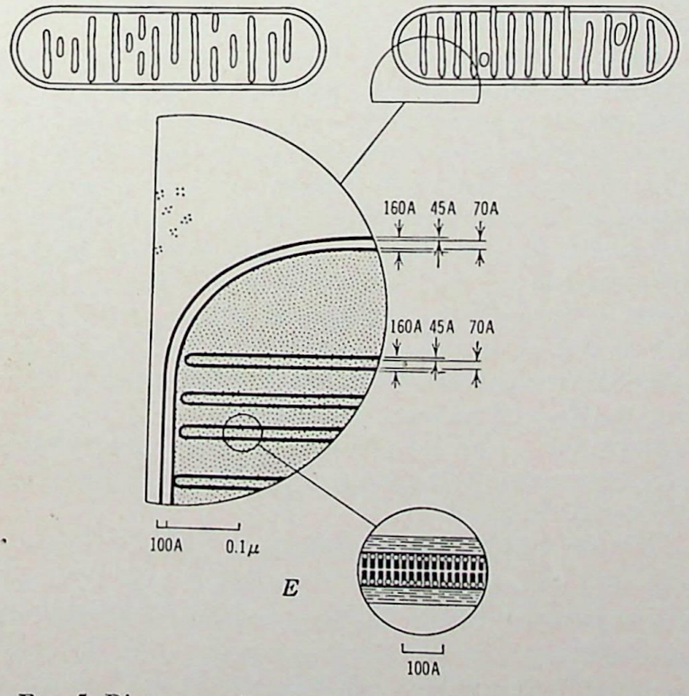


FIG. 5. Diagrammatic representation of the structure of mitochondrion according to Sjöstrand.[33] Note the double-layer character of outer membrane and the transverse "cristae." The magnified cross-section of the membrane suggests two protein monolayers separated by an oriented double layer of lipid molecules [from F. S. Sjöstrand in *Fine Structure of Cells—Symposium, Eighth Congress on Cellular Biology, Leyden, 1954* (Interscience Publishers, Inc., New York, 1955), pp. 16, 222].

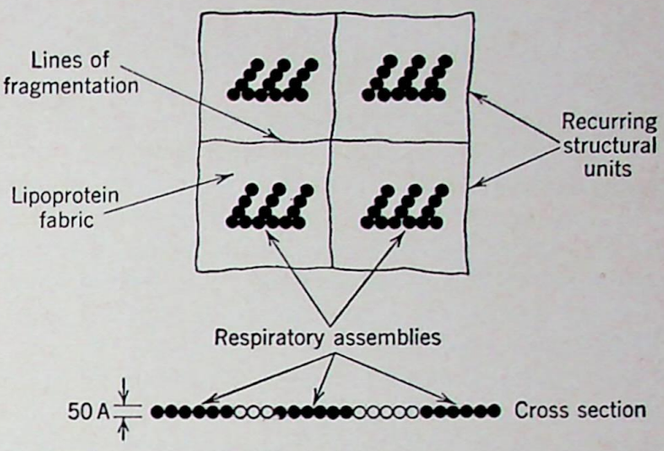


FIG. 6. Respiratory-enzyme assemblies in recurring units of mitochondrial membrane. The assembly consists of 6 electron-carrier protein molecules, supplemented by three sets of coupling enzymes, each consisting of three proteins as an approximation. This is a simplified representation since there may be several flavoprotein molecules per mole of cytochrome-*a* and several DPN-dehydrogenase molecules[1] [from A. L. Lehninger, C. L. Wadkins, C. Cooper, T. M. Devlin, and J. L. Gamble, Jr., Science 128, 450 (1958)].

The structural details of mitochondria have been visualized by electron microscopy in thin sections of intact tissues in the laboratories of Sjöstrand, Palade, and others.[33] Although there are some differences in interpretation of electron micrographs of these organelles, the diagrammatic representation of Sjöstrand (Fig. 5) shows that the mitochondrion is surrounded by an outer double-layered membrane and also contains double-layered lamellar structures across the lumen called the "cristae" by Palade. These membranes, which are easily separable into fragmented form, are rich in lipid, containing some 30 to 40% which is mostly phospholipid. Sjöstrand has pointed out, as shown in his diagram, that the double-layered membranes correspond approximately to two monolayers of protein molecules separated by a double layer of oriented lipid molecules in the fashion suggested, and a similar arrangement in the cristae. These dimensions are constant in the mitochondria of all cell types examined.[33] The inner matrix of the mitochondrion contains soluble proteins and enzymes, electrolytes, and nucleotides which are released rather easily into the medium by mechanical or by osmotic rupture of the membranes. The results of many investigators suggest that most of the substrate-level enzymes of the Krebs cycle and the fatty-acid oxidation cycle are present either in the lumen or in the intercristal space. They are released easily in soluble form after damage of the membrane. On the other hand, it has been a universal finding that the enzymes involved in electron transport and coupled oxidative phosphorylation are present entirely in the insoluble membranes in organized "solid-state" systems; the separate enzymes of this system are not dissociated easily in truly soluble form.[5,16,34–36]

In phosphorylating subfragments of the mitochondrial membranes, the several electron carriers making

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

up the respiratory chain occur in approximately equimolar ratios,[29] suggesting that each respiratory "assembly" consists of the six (or more) carrier molecules arranged in contiguous manner, presumably adjacent to each other. These are supplemented at three points along the chain by three sets of two or three (or more) auxiliary enzyme proteins which are necessary for the energy-coupling process, as in the coupling mechanism postulated. Such an arrangement thus would form a complete respiratory-phosphorylating multienzyme assembly, as is shown in Fig. 6.

The question arises as to the distribution of such respiratory assemblies along the membrane. We have subjected sonically-prepared fragments of the mitochondrial membrane to differential ultracentrifugation and have obtained a wide spectrum of particle sizes. The relative distribution of respiratory carriers in these particles of widely different sedimentation rate has been examined with the finding that, no matter what the size of the fragment, the ratio of the respiratory carriers to total protein was always constant, and the ratio of carriers to each other was relatively constant also. Phosphorylation activity is preserved in all fractions.[30] This finding indicates that the respiratory carriers are distributed more or less evenly along the membrane, and that fragmentation of the membrane produces pieces of different size made up of multiples of a basically recurring unit, as is indicated in the diagram, where each recurring unit contains a complete and organized respiratory assembly.[5]

Calculations involving the known extinction coefficients of some of the respiratory carriers such as cytochrome-*c* indicate that, in the membranes of a single rat-liver mitochondrion, there may be several thousand such respiratory assemblies. Still further analyses of this kind make it possible to estimate the total fraction of the mitochondrial membrane substance contributed by catalytically active proteins of these assemblies (assuming twelve protein molecules per assembly, each having a molecular weight of 10^6) to be as high as 20% and possibly higher. From such considerations, it can be calculated that the average particle weight of a single respiratory assembly, together with its "filler" protein and lipid, may be 7×10^{-7} approximately.[37]

From such considerations, it is clear that the mitochondrial membrane is not simply a "dead wall," but rather a highly complex fabric in which are embedded the respiratory assemblies in a highly oriented and purposeful manner. These respiratory assemblies thus represent the very warp and woof of the membrane fabric.

MECHANICAL PROPERTIES OF MITOCHONDRIAL MEMBRANE

The mitochondrial membrane is not only the site of the highly organized respiratory assemblies, but it is also a structure capable of characteristic and selective permeability and active-transport phenomena. By optical measurements, it has been observed that the permeability may be altered by the ambient concentrations of ATP, magnesium, sulfhydryl reagents, and inorganic phosphate. In addition, we have found that the permeability of the mitochondrial membrane may be changed very drastically by addition of minute concentrations of the thyroid hormone, which causes rapid swelling and increased permeability to sucrose and other small molecules.[38–40] The mitochondrial membrane is capable of quite drastic extension and swelling under certain circumstances; it may increase in volume as much as several fold without actual rupture of the membranes (cf. Werkheiser and Bartley[41]). Phase contrast films of mitochondria in an intact cell show that they have remarkable plasticity of shape, and undergo rhythmic swelling and contraction as they wander through the canaliculi of the cytoplasm. The contractile properties of the mitochondrial membrane also can be observed *in vitro* when adenosine triphosphate is added to suspensions of heart mitochondria or when phosphorylating respiration is instituted. These agencies cause a rapid contraction of the mitochondrial membrane with the extrusion of water, and it has been postulated that the mitochondria, at least in specialized tissue such as the kidney, are active in water transport.[16,42] The contractility of the mitochondrial membrane thus may resemble the configurational changes in the myosin molecule similarly coupled to ATP energy. The number of water molecules extruded from mitochondria is out of all proportion to the number of molecules of ATP formed or utilized.

Very recently, we found that the permeability and contractility of the mitochondrial membrane are conditioned by the oxidation state of the respiratory carrier enzymes present in the membrane. Thyroxine and inorganic phosphate, which are potent swelling agents of mitochondria, increasing the permeability of the membrane, have this action only when the respiratory carriers are in the *oxidized* state. The action of these compounds results in a relaxation of the membrane and in increased permeability to certain test substances such as sucrose. On the other hand, when the respiratory carriers are maintained in the fully *reduced* state, the permeability of the mitochondrial membrane cannot be changed by addition of phosphate or thyroid hormone.[39,40]

The mitochondrial respiratory assemblies thus not only are the site of phosphorylating oxidation and major building blocks of the membrane, but also are controlling factors of rather dramatic mechanical and configurational changes which govern the rate of entry of essential metabolites such as substrates and inorganic phosphate, as well as the exit of respiration products such as H_2O, ATP, carbon dioxide, etc. The membrane-bound respiratory assembly thus becomes an example of a "mechanoenzyme" system, to use

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

Engelhardt's term for myosin. Indeed, the mitochondrial membrane has many characteristics resembling those of myosin, such as ATP-ase activity, DNP-stimulation, striking ionic strength relationships, as well as the configurational changes.

SOME BIOPHYSICAL PROBLEMS ASSOCIATED WITH RESPIRATORY-ENERGY CONVERSIONS AND "SOLID-STATE ENZYMOLOGY"

Most well-known enzyme systems, such as the citric-acid cycle and glycolytic cycle, for example, involve low molecular-weight, rapidly diffusible intermediate molecules like di- and tri-carboxylic acids or phosphorylated sugars, as shuttles between relatively slowly diffusing enzyme proteins in an alternating manner. The respiratory-chain system, however, differs strikingly; it involves protein-protein interactions exclusively without low molecular-weight intermediates (Fig. 7). Since the rate of diffusion of protein molecules is extremely low as compared with that of simple sugars, amino acids, etc., a multienzyme sequence of a dozen consecutive reactions, each of which is brought about through protein-protein collisions alone, can be expected to occur at very low rates only if there were no means to direct collisions. In fact, it can be calculated readily that the known high rates of respiration of cells could not occur if all of the enzymes involved were distributed equally in free solution within the volume of the cell.[43] The assembly of the individual proteins making up this system into a solid-state array, as in the mitochondrial membrane, provides a molecular adaptation to overcome this kinetic problem. A variety of questions arises, however.

For one thing, do ordinary mass-law and ideal-collision considerations govern the behavior of protein molecules in such a "solid-state" array? The question is important and underlies all attempts to analyze reaction kinetics and reaction mechanisms in solid-state enzyme systems by the application of Michaelis-Menten principles and other kinetic parameters. Chance has commented on this problem.[1]

CONSECUTIVE REACTION PATTERNS IN MULTIENZYME SYSTEMS

I. Low molecular-weight intermediates as "shuttles":

$$S \rightleftarrows (P_1) \rightleftarrows S \rightleftarrows (P_2) \rightleftarrows S \rightleftarrows (P_3) \rightleftarrows S \rightleftarrows (P_4) \text{ etc.}$$

Example: Glycolysis via phosphorylated intermediates.

II. Protein-protein interactions; no "shuttles":

$$(P_1) \rightleftarrows (P_2) \rightleftarrows (P_3) \rightleftarrows (P_4) \text{ etc.}$$

Example: Electron transport

FIG. 7. Diffusion rates as factors in molecular interactions in multienzyme systems. Low molecular-weight substrate (product) molecules S, S_x, etc., have high-diffusion coefficients and may "shuttle" between high molecular-weight enzyme proteins; in the respiratory chain, slow-moving proteins must interact.

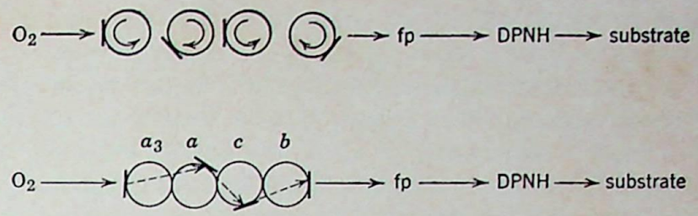


FIG. 8. Modes of electron transfer along chain of fixed electron-carrier molecules (after Chance and Williams[1]).

New methods thus will be required to examine the interaction of the individual catalyst molecules in such solid-state arrays, and to determine the translational, rotational, and vibrational degrees of freedom involved in these protein-protein interactions. Possibly, useful information will come from refined application of both electron-paramagnetic and nuclear-magnetic resonance spectroscopy, to supplement the already-extensive light spectroscopic studies of Chance and his colleagues.[1] Another question arises. How can the mechanism of electron transfer along the respiratory chain be accounted for, in view of the known dimensions of some of these proteins and their prosthetic groups, even assuming close juxtaposition of carriers? Chance has shown[1] two alternatives for the mechanism of electron transfer from one protein carrier to another along the chain (Fig. 8). In one, restricted rotation is postulated to permit collision of the prosthetic groups. In the second, the molecules are fixed, and electrons then must pass through the protein moieties to the prosthetic groups, possibly by π-electron interactions as in the so-called "charge-transfer" complexes. The passage of electrons or energy through the protein moiety of heme proteins has been considered frequently, since the classical experiments of Bücher on the constant quantum efficiency of 1.0 in the photodissociation of carbon monoxide myoglobin at all wavelengths tested.[44] Recently, Shore and Pardee have considered the physics of such energy transfers, which may take place through fluorescence and sensitization phenomena.[45,46] All of the respiratory carriers have quite characteristic fluorescence spectra. Szent-Györgyi, Weber, and others have suggested that fluorescence phenomena may conceivably play a physiological role in energy transfers along the carrier chain.[47]

The occurrence of such solid-state arrays of catalytically active proteins, the individual members of which are so difficult and as yet impossible to separate from each other with preservation of activity, presents a whole new area of protein chemistry and techniques. To date, it has been sufficiently difficult to examine the physical parameters of proteins in more or less ideal solutions; the interaction of two proteins with each other in dilute solution actually has been just barely touched upon in recent years. However, these far more complex solid-state arrays described above obviously present difficulties of a much greater order of magnitude. Similarly, the separation and identification of

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

such proteins making up solid-state protein arrays will involve a whole new category of experimental approaches, as students of this area of enzyme chemistry already are painfully aware. If new techniques finally will permit separation of all of the carrier enzymes in soluble form, then reconstruction of these systems *in vitro* also can be expected to present great difficulty, because of the problem of restoring the spatial relationships among the members of such a complex in the test tube. Perhaps full *in vitro* reconstruction to yield soluble systems of high-catalytic activity never really will be attained.

These are but a few of the many biophysical and molecular problems posed by the solid-state character of the respiratory-energy transformers of mitochondria.

BIBLIOGRAPHY

[1] B. Chance and G. R. Williams, Advances in Enzymol. **17**, 65 (1956).

[2] F. E. Hunter, Jr., in *Phosphorous Metabolism*, W. D. McElroy and B. Glass, editors (The Johns Hopkins Press, Baltimore, Maryland, 1951), Vol. I, p. 297.

[3] A. L. Lehninger in *Phosphorous Metabolism*, W. D. McElroy and B. Glass, editors (The Johns Hopkins Press, Baltimore, Maryland, 1951), Vol. 1, p. 344.

[4] A. L. Lehninger in *Harvey Lectures* (Academic Press, Inc., New York, 1953–1954), Vol. IL, p. 176.

[5] A. L. Lehninger, C. L. Wadkins, C. Cooper, T. M. Devlin, and J. L. Gamble, Jr., Science **128**, 450 (1958).

[6] H. M. Kalckar, Chem. Revs. **28**, 71 (1941).

[7] N. O. Kaplan in *The Enzymes*, J. B. Sumner and K. Myrbäck, editors (Academic Press, Inc., New York, 1951), Vol. II, p. 55.

[8] F. Lipmann, Advances in Enzymol. **1**, 99 (1941).

[9] K. Burton, Nature **181**, 1594 (1958).

[10] T. L. Hill and M. F. Morales, J. Am. Chem. Soc. **73**, 1656 (1951).

[11] T. Benzinger and C. Kitzinger, Z. Naturforsch. **10b**, 375 (1955).

[12] K. Burton and H. A. Krebs, Biochem. J. **54**, 94 (1953).

[13] K. Burton, Biochem. J. **59**, 44 (1955).

[14] M. J. Johnson in *Respiratory Enzymes*, H. A. Lardy, editor (Burgess Publishing Company, Minneapolis, 1949).

[15] E. P. Kennedy and A. L. Lehninger, J. Biol. Chem. **172**, 847 (1948).

[16] L. Ernster and O. Lindberg, Ann. Rev. Physiol. **20**, 13 (1958).

[17] D. E. Green, Symp. Soc. Exptl. Biol. **10**, 30 (1957).

[18] C. L. Wadkins and A. L. Lehninger, J. Biol. Chem. (to be published).

[19] M. Friedkin and A. L. Lehninger, J. Biol. Chem. **178**, 611 (1949).

[20] A. L. Lehninger, J. Biol. Chem. **190**, 345 (1951).

[21] S. O. Nielsen and A. L. Lehninger, J. Am. Chem. Soc. **76**, 3860 (1954).

[22] C. Cooper and A. L. Lehninger, J. Biol. Chem. **224**, 561 (1957).

[23] E. C. Slater, Nature **172**, 975 (1953).

[24] C. L. Wadkins and A. L. Lehninger, J. Biol. Chem. (to be published).

[25] C. Cooper, T. M. Devlin, and A. L. Lehninger, Biochim. et Biophys. Acta **18**, 159 (1955).

[26] C. Cooper and A. L. Lehninger, J. Biol. Chem. **219**, 489 (1956).

[27] See reference 26, p. 519.

[28] C. Cooper and A. L. Lehninger, J. Biol. Chem. **224**, 547 (1957).

[29] T. M. Devlin, J. Biol. Chem. (to be published).

[30] T. M. Devlin and A. L. Lehninger, J. Biol. Chem. **219**, 507 (1956).

[31] T. M. Devlin and A. L. Lehninger, J. Biol. Chem. (to be published).

[32] C. L. Wadkins and A. L. Lehninger, J. Am. Chem. Soc. **79**, 1010 (1957).

[33] F. S. Sjöstrand in *Fine Structure of Cells—Symposium, Eighth Congress on Cellular Biology, Leyden, 1954* (Interscience Publishers, Inc., New York, 1955), pp. 16, 222.

[34] E. G. Ball and R. J. Barrnett, J. Biophys. Biochem. Cytol. **3**, 1023 (1957).

[35] D. E. Green in *Harvey Lectures* (Academic Press, Inc., New York, 1956–1957), Vol. LII, p. 177.

[36] D. M. Ziegler, A. W. Linnane, D. E. Green, C. M. S. Dass, and H. Ris, Biochim. et Biophys. Acta **28**, 524 (1958).

[37] A. L. Lehninger in *Recent Advances in Molecular Biology* (to be published).

[38] A. L. Lehninger in *Enzymes: Units of Biological Structure and Function*, O. H. Gaebler, editor (Academic Press, Inc., New York, 1956), p. 217.

[39] A. L. Lehninger and B. L. Ray, Biochim. et Biophys. Acta **26**, 643 (1957).

[40] A. L. Lehninger, B. L. Ray, and M. Schneider, J. Biophys. Biochem. Cytol. (to be published).

[41] W. C. Werkheiser and W. Bartley, Biochem. J. **66**, 79 (1957).

[42] C. A. Price, A. Fonnesu, and R. E. Davies, Biochem. J. **64**, 754 (1956).

[43] A. G. Ogston and O. Smithies, Physiol. Rev. **28**, 283 (1948).

[44] T. S. Bücher, Advances in Enzymol. **14**, 1 (1953).

[45] V. G. Shore and A. B. Pardee, Arch. Biochem. Biophys. **60**, 100 (1956).

[46] V. G. Shore and A. B. Pardee, Arch. Biochem. Biophys. **62**, 355 (1956).

[47] G. Weber, Biochem. J. **51**, 155 (1952).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

17

Energy Reception and Transfer in Photosynthesis*

MELVIN CALVIN

Department of Chemistry, University of California, Berkeley 4, California

THE article by Lehninger (p. 136) presents a detailed and excellent description of how a cell can obtain energy by the combustion of carbohydrate. This article describes the reverse process, namely, how the green cells of plants are able to transform electromagnetic energy into chemical energy—by the absorption of carbon dioxide and water, which are the end products of the animal cell, and by the absorption of light—and how they produce the foodstuffs which are the beginning of the process of combustion. Figure 1 illustrates diagrammatically the content of this article.

The starting points in this case are carbon dioxide and water which contain the elements carbon, hydrogen, and oxygen in their lowest energy forms with respect to biological processes. The chemical energy which is accumulated is represented here (Fig. 1) in the form of oxygen (molecular oxygen), on the one hand, and a carbohydrate, on the other. The process itself has been divided, both theoretically and physically, into two rather easily separable stages. The first of these is the absorption of light by chlorophyll or by some related pigments and the subsequent separation of water into a reducing agent, here represented by [H], and some oxidizing fragment not specifically designated here but presumably one of the *A*, *B*, *C*, series. The oxidizing agent, or the primary oxidant, ultimately becomes molecular oxygen. In the second stage, the reducing agent is used to reduce carbon dioxide to the level of carbohydrate and other plant materials.

In order to see how the energy of light actually is accumulated in chemical form, it seems wise to describe what is known about the sequence from carbon dioxide to carbohydrate, and to determine at what point in that sequence the energy ultimately derived from the light enters, and from that point on to recognize and to define the problem of the primary quantum conversion into its first recognizable chemical form. Consider first in some detail what is known about the path of carbon so that one can define more precisely into what sort of energy the light must be converted in order to carry out that process.

With the availability of radiocarbon (carbon-14) from the nuclear reactors some 15 years ago, it became possible for our laboratory to trace this sequence in some detail. The plant material used in most of the experiments was the unicellular green alga, *Chlorella*, and occasionally the alga, *Scenedesmus*; higher plants as well as separated photosynthetic material were used also. Figure 2 shows a photomicrograph of the algae cells commonly used; these are the *Chlorella* cells and the green stuff contained in a cup-shaped chloroplast can be seen. It is illustrated well by one of the cells in the upper right-hand corner.

The steps taken to trace the carbon sequence are as follows. The first operation constitutes a selection of cultures, which are grown in 200-cc flasks, and are transferred later into much larger continuous one-liter culture flasks. These are called shake-flask cultures in which algae can be maintained for years at a time.

The most recent type of culture device that is used in our laboratory is a continuous tube culture in which the density of the cells is monitored by a photoelectric cell which controls the automatic feeding of the medium, so that the cells are maintained in a steady state of growth.

The algal sample then is harvested and is used for the feeding of radiocarbon which is done in a special "hot box." In this box, the cells are placed in a little vessel (lollipop) between lights and are adapted with the concentration of normal carbon dioxide of interest. Radioactive carbon dioxide then is administered to the

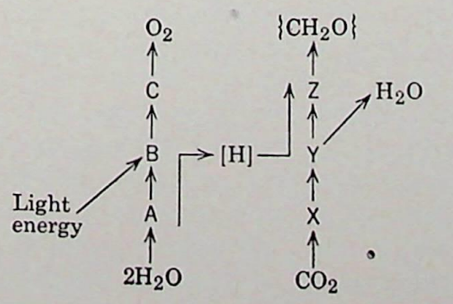


FIG. 1. Elementary photosynthesis scheme.

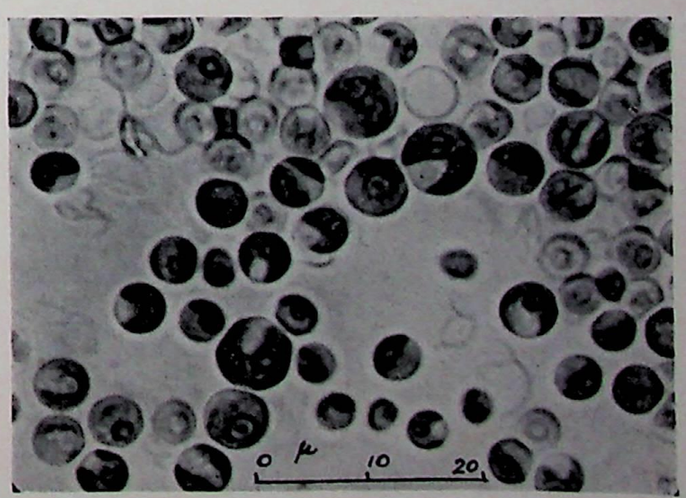


FIG. 2. Photomicrograph of *Chlorella* cells.

* The work described herein was sponsored by the U. S. Atomic Energy Commission.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

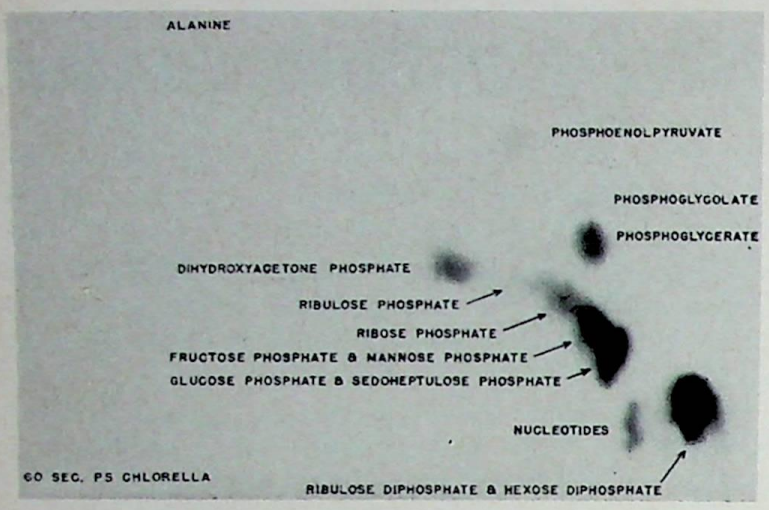


FIG. 3. Chromatogram of extract from algae, indicating uptake of radiocarbon during photosynthesis (60 sec).

adapted cells for a suitable length of time in order to trace the paths taken by the carbon atoms. The radioactive carbon usually is injected in the form of a solution of sodium bicarbonate. It is kept in contact with the cells for a specified period of time after which the cells are killed by a variety of methods; for example, the cells may be dropped into methanol at room temperature. The cell extract then is analyzed by the method of paper chromatography for the radioactive compounds which it may contain. In order to achieve this analysis, the extracts must be concentrated and a vacuum evaporator is used in a routine fashion to reduce the volume from 200 cc, or a liter, down to a cubic centimeter or so.

From this concentrated extract, an aliquot is taken and placed on the corner of a piece of filter paper for chromatographic separation. Prior to chromatography, the radioactivity of the origin is counted in a quantitative way. The filter paper then is hung in a box, in a trough in which a solvent is placed which passes over the filter paper and spreads the compounds down the side of the paper according to their relative solubilities in the solvent. The most soluble run the most rapidly. This procedure results in a set of spots along the side of the filter paper, depending upon the properties of the compound being analyzed. Some of these compounds overlap each other in one solvent system. The paper

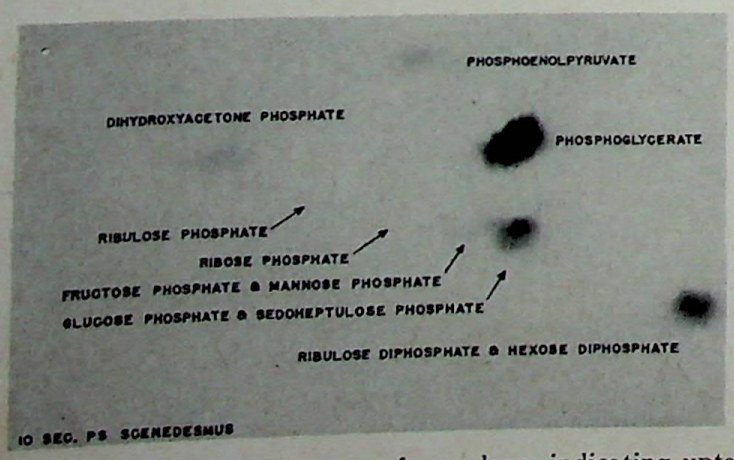


FIG. 4. Chromatogram of extract from algae, indicating uptake of radiocarbon during photosynthesis (10 sec).

then is removed from the box and dried overnight. The paper is rotated 90° and placed in another trough, in another box, with another solvent. The starting point is now a whole series of spots along the top edge of the paper. Another solvent is put in the trough and it spreads those spots out again in a similar operation. After this operation is completed, the paper is dried for a second time. The next problem is to locate the radioactive materials on the paper.

The coordinates of the material with respect to the origin constitute a physical property which is useful in the identification of the compound being analyzed. The compounds are not colored, and the only property that can be used to locate them is their radioactivity. This is done by placing the paper in contact with a sheet of photographic single-coat blue-sensitive x-ray film which becomes exposed by the radioactivity on the paper. Wherever there is a radioactive spot on the paper, there appears an exposed area on the film after a suitable period of time. For quantitative work, the film, covered

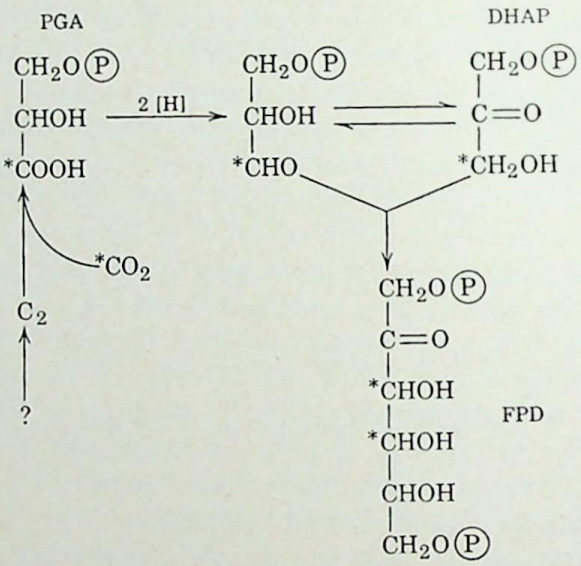


FIG. 5. Path of carbon from CO_2 to hexose during photosynthesis.

by the paper, is placed on an x-ray viewer. The paper is explored with a Geiger counter. The compound can be identified then by its coordinates, and its amount can be determined by the amount of radioactivity found in the spot. A greater or lesser degree of resolution depends upon the nature of the solvent systems used and upon the time used for the chromatography.

Figure 3 shows a chromatograph picture of the extract of a 60-sec illumination of *Chlorella*. One can see that a 60-sec illumination is much too long to find the earliest compounds into which the carbon enters in the course of its conversion from carbon dioxide to carbohydrate. Figure 4 shows a chromatogram of a shorter illumination (10 sec). Here, one compound, phosphoglyceric acid, dominates the scene. Our laboratory has been able to get the same type of sequence of events with isolated chloroplasts plus a number of cofactors. The phosphoglycerate appears no matter how the chloroplasts or the algae are killed: whether they are

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

killed hot or cold, whether they are killed in alcohol, acetone, etc.

This, then, is an initial clue that the first isolable stable compound obtainable by these methods, or at least identifiable by these methods, is phosphoglyceric acid:

$$\begin{array}{l} CH_2\text{—}CHOH\text{—}CO_2H \\ | \\ OPO_3H \end{array} ,$$

a three-carbon compound containing a low-energy phosphate group.

The next problem is to determine which of these carbon atoms is radioactive. This has been done by chemical-degradation methods. By taking the compound (phosphoglyceric acid) apart, one carbon atom at a time, it was found that the carboxyl group became radioactive first and the other two later. From this, together with the degradation of sugar molecules that came out in the same experiment, our laboratory was able to determine how the sugar molecule was constructed.

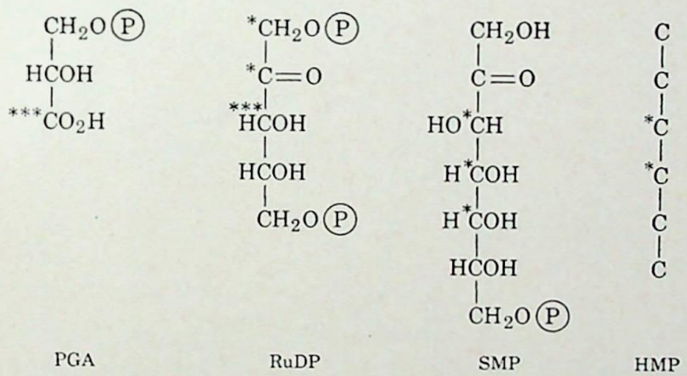


FIG. 6. Distribution of radioactive carbon in certain sugars.

Figure 5 shows what was supposed to have occurred. The phosphoglyceric acid is shown as PGA. By reduction, this goes to triose phosphate. If the ketose phosphate is isomerized, and then combined with the isomer, a hexose diphosphate can be formed with the radioactive carbon atoms in the middle of the molecule. In this manner, the six-carbon molecule can be formed, but one does not know the origin of the three-carbon compound. Although two-carbon-containing molecules have not been found, Fig. 6 illustrates some findings of our laboratory.

In addition to the PGA, there is a five-carbon-atom compound, a sugar (ribulose diphosphate), a seven-carbon-atom sugar (sedoheptulose diphosphate), and, of course, the six-carbon-atom sugars. The stars on Fig. 6 indicate some idea of the way in which the radioactivity is distributed in these various sugar compounds. The heptose and the pentose can be made from the hexose, as shown in Figs. 7 and 8.

Figure 7 illustrates the method by which the heptose may be produced. From one molecule of hexose and one

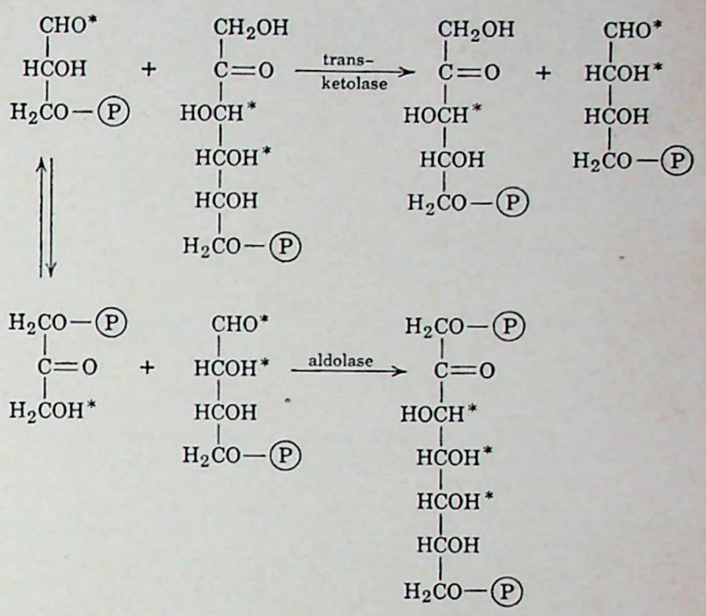


FIG. 7. Formation of a heptose from triose and hexose.

molecule of triose (taking off the top two carbon atoms of the hexose), a pentose and tetrose can be formed; the tetrose is labeled in the top two carbon atoms. The tetrose then can combine with a triose to make a heptose with the proper distribution of radioactive carbon.

Figure 8 shows the way in which the pentose is put together, by combination of a heptose and a triose, in the same kind of reaction (the transketolase reaction) leading to two different pentoses which are in equilibrium with each other. This analysis does not distinguish between the two pentoses. All of these rearrangements are done at the sugar level; triose, tetrose, pentose, hexose, and heptose are all at the same redox level. They are all of them very nearly at the same energy level and there is thus practically no energy required for these rearrangements. However, no experiments of this type gave the desired information—namely, the origin of the three-carbon piece in the first place. This awaited a quite different kind of experiment, an experiment in which a steady state first was established in the organism, after which some environmental variable was changed suddenly. The transients that resulted from changing some of these variables were examined.

Figure 9 shows the results of such an experiment. A steady state is established by feeding the radiocarbon

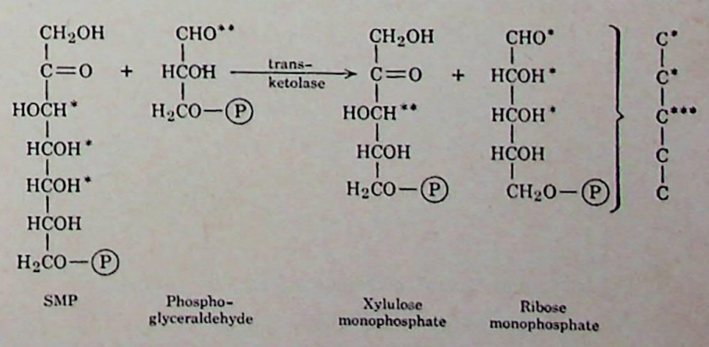


FIG. 8. Proposed scheme for labeling of pentose.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

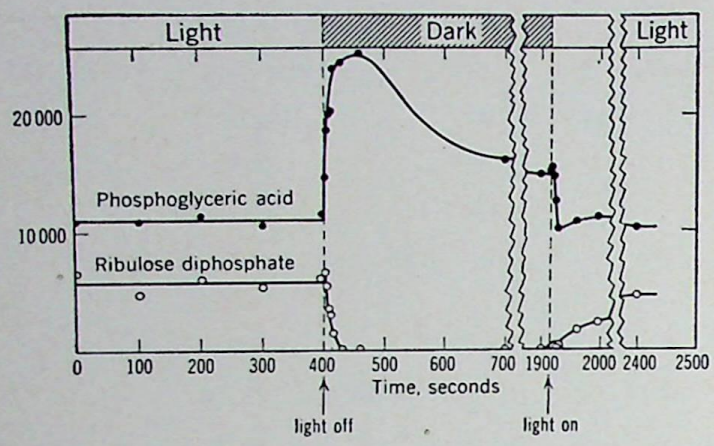


FIG. 9. Light-dark changes in concentrations of phosphoglyceric acid and ribulose diphosphate.

long enough to the plant to saturate the phosphoglyceric acid and the other compound mentioned. The lights are turned off suddenly, and the transient ensues. The phosphoglyceric acid rises suddenly and the ribulose diphosphate falls precipitously. This complementary behavior is the clue needed for the relationship between ribulose diphosphate and phosphoglyceric acid. It seemed as though the ribulose diphosphate was disappearing by combining with the carbon dioxide (that is, five carbons plus one, making a total of six carbon atoms) to produce two molecules of phosphoglyceric acid. If this is the case, then the relationships of the various compounds can be shown diagrammatically, as in Fig. 10.

In Fig. 10, the ribulose diphosphate combines with carbon dioxide to form phosphoglyceric acid which is reduced by light to triose, triose then going through this series of sugar rearrangements (shown in Figs. 6, 7, and 8) back again to the pentose. Turning the light off stops this reaction. When the reduction reaction is stopped, phosphoglyceric acid builds up and the ribulose diphosphate disappears. Figure 10 simply expresses in a scheme what the transient experiment revealed.

But this scheme (Fig. 10) predicts another type of transient. If the light is kept on and the CO_2 stopped, there should be a different kind of transient; namely, the ribulose diphosphate should build up suddenly and the amount of the phosphoglyceric acid should fall. This experiment has been accomplished with considerable difficulty. The results are shown in Fig. 11.

Figure 11 shows the steady state for ribulose diphosphate and the steady state for phosphoglyceric acid with the CO_2 at a concentration of 1%. At the vertical line, the CO_2 concentration is shifted from 1 to 0.003% by turning

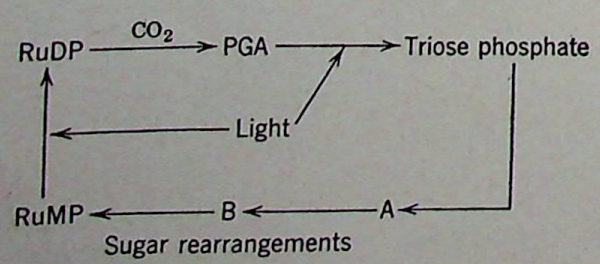


FIG. 10. Formation of PGA from RuDP.

the stopcocks. Under those circumstances, the predicted changes were observed, at least in the initial phase of the transient. The amount of phosphoglyceric acid fell and the amount of ribulose diphosphate rose. There is a number of kinetic oscillations here which are reminiscent of the kinds of oscillations one gets in circuitry, and possibly they are analogous. One or two attempts were made to reproduce these oscillations by putting first-order rate constants into the various reactions that are involved here and then running them through a digital computer. This kind of oscillation can be obtained but this work has not been pursued yet beyond the elementary stage of the first kind of transient. This kind of study will lead to much more-detailed knowledge of the mechanism of cellular response to changes in external or internal environments. It is a very simple system to use and one which is amenable to complete analysis, both experimentally in terms of the compounds

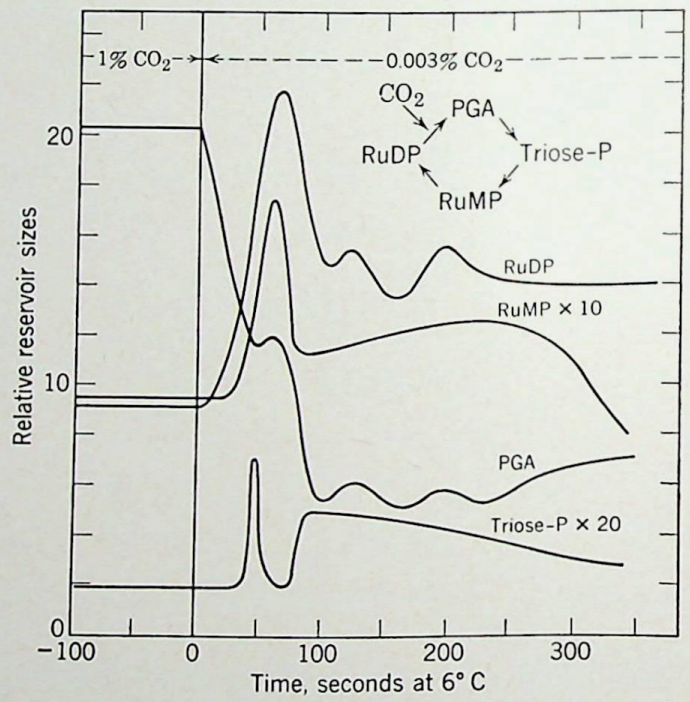


FIG. 11. Transients in the regenerative cycle.

involved and theoretically in terms of the simple kinetics involved.

Figure 12 shows the completed photosynthetic cycle in which there are put together all of the rearrangements of hexose and triose through heptose and pentose, back again to the ribulose diphosphate which then picks up carbon dioxide to make two molecules of phosphoglyceric acid.

In trying to visualize this particular step, a proposed mechanism for this reaction is shown in Fig. 13. Here, the ribulose diphosphate is written as the ene-diol, combining with bicarbonate ion to form an intermediate, hypothetical up to this point, an α-hydroxy-β-keto acid, which then is hydrolyzed to give two molecules of phosphoglyceric acid. This α-hydroxy-β-keto acid, according to our chemical knowledge, would be very unstable either to decarboxylation, in which case it would

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

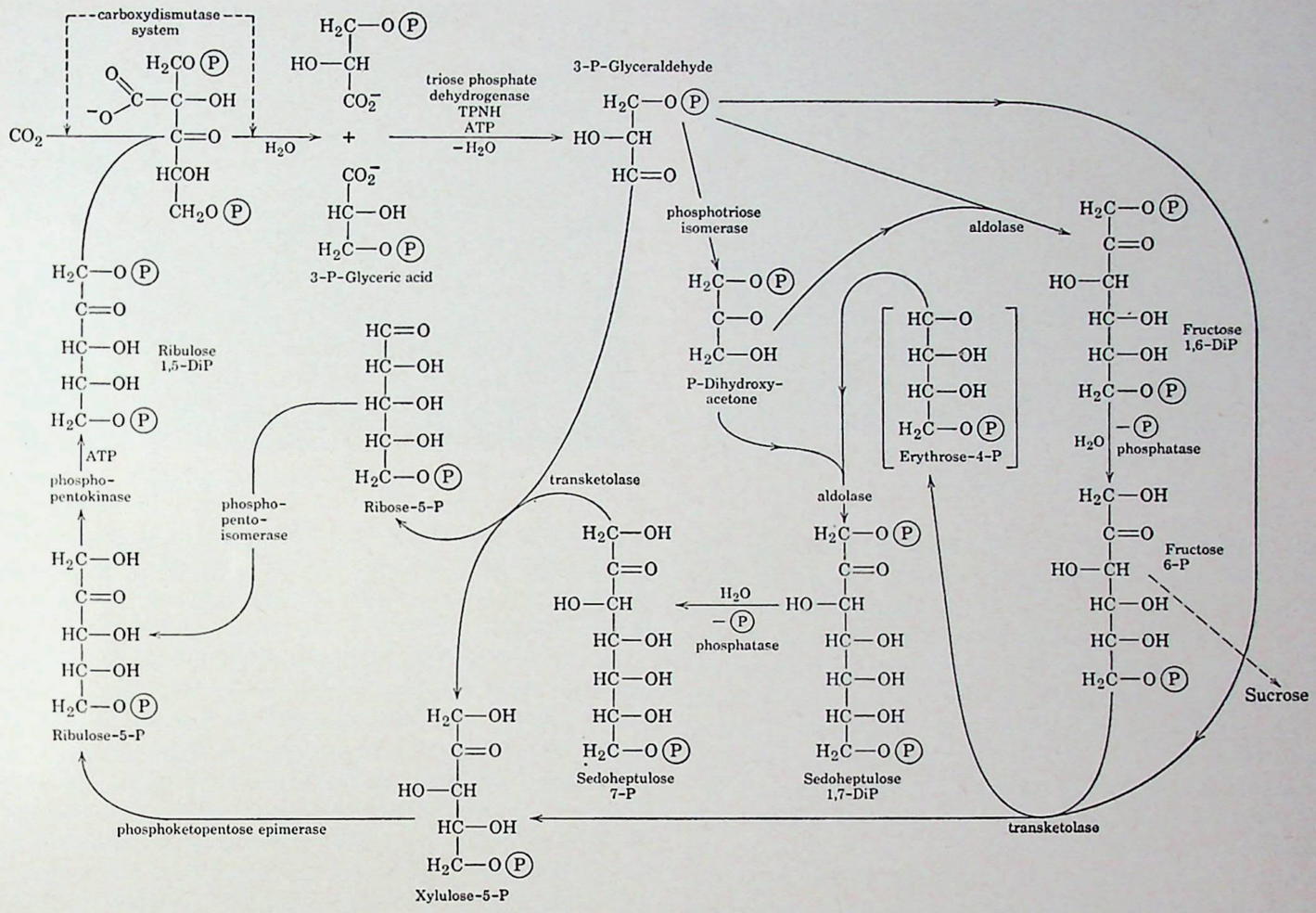


FIG. 12. The photosynthetic carbon cycle.

lead back to ribulose diphosphate, or to hydrolysis, in which case it would lead the other way. Fortunately, in our laboratory, a close relative of this intermediate, and probably some of the compound itself, has been found.[1] On theoretical grounds, it was decided that one might expect a compound of this sort to appear down in the diphosphate area of the chromatogram. Figure 14 shows that chromatogram which ran in a solvent for 48 hr in both coordinates. What was originally a single spot, which is dominantly ribulose diphosphate, now breaks up into at least three spots. The principal spot is the ribulose diphosphate; another one is hexose diphosphate and heptose diphosphate; and the last spot turned out to be a keto-acid diphosphate. It is not the β-keto-acid but rather the γ-keto-acid diphosphate which apparently is an artifact of the method of killing, but it does come from the β-keto-acid diphosphate which is still a labeled compound and does show its presence in small amounts.

In Fig. 15, one sees the diphosphate plus some dephosphorylated compounds, particularly the γ-keto-acid

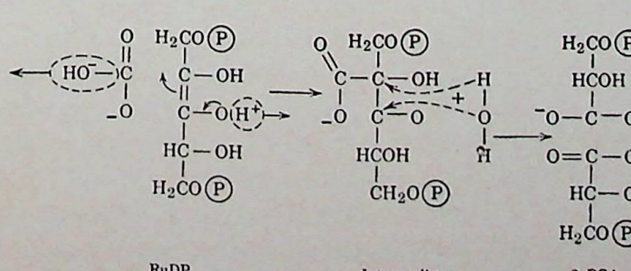


FIG. 13. Mechanism of carboxylation reaction.

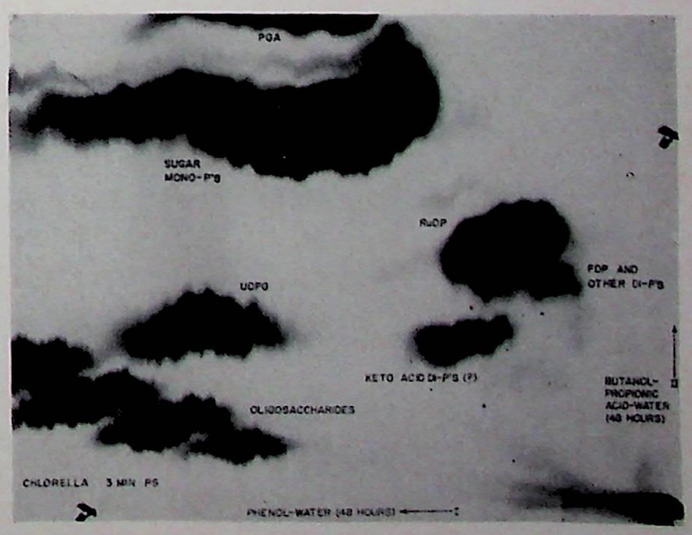


FIG. 14. Chromatogram of extract of *Chorella* after three minutes of photosynthesis in the presence of radiocarbon. The solvents were allowed to run for 48 hr in each dimension.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

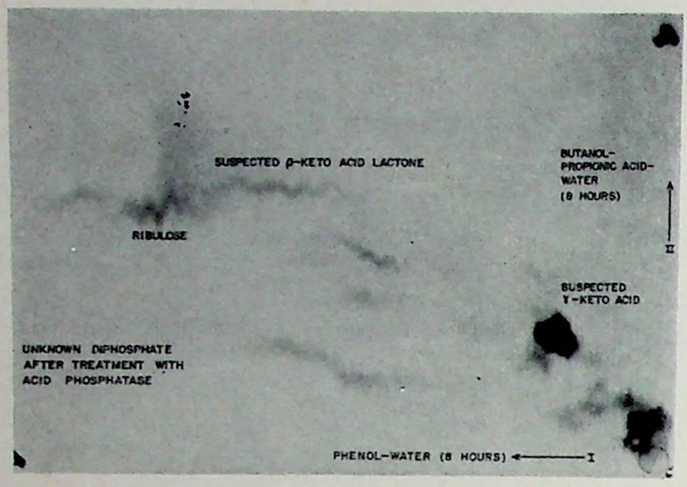


FIG. 15. Unknown sugar phosphate after treatment with acid phosphatase.

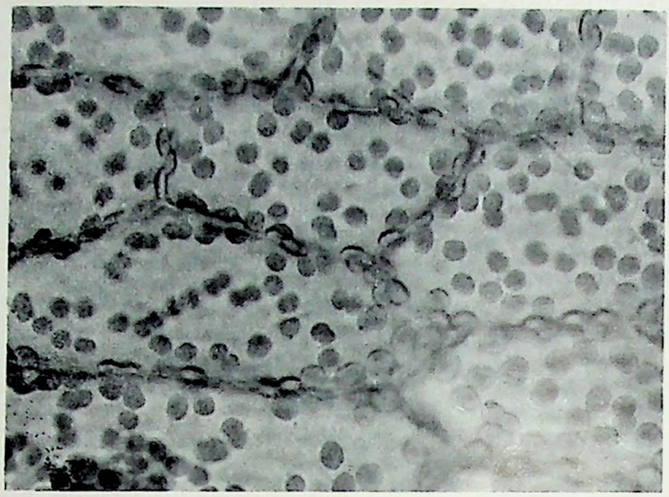


FIG. 17. Photomicrograph of chloroplasts from liverwort.

diphosphate. And here is seen a trace of the β-keto acid in its lactone form because it is stabilized as a lactone, enough to catch it on a chromatogram. Racker[2] has carried out this whole sequence (Fig. 12) by collecting all of the enzymes that were indicated in that figure. By putting in the suitable substrates, ribulose and carbon dioxide, he was able to pull out glucose phosphate from the CO_2. The immediate sources of energy are the two compounds adenosine triphosphate (ATP) and reduced pyridine nucleotide (TPNH). Figure 16 shows the needed relationships.

Figure 16 shows the photosynthetic carbon cycle in a simplified form. Carbon dioxide enters to make the β-keto acid which then goes to the phosphoglyceric acid. The only points of entry of energy into this system, as it is written now, are where there is a need for ATP and for reduced pyridine nucleotide. (These are the points of entry of energy into this wheel.) These points are the gears which drive the cycle in a forward direction.

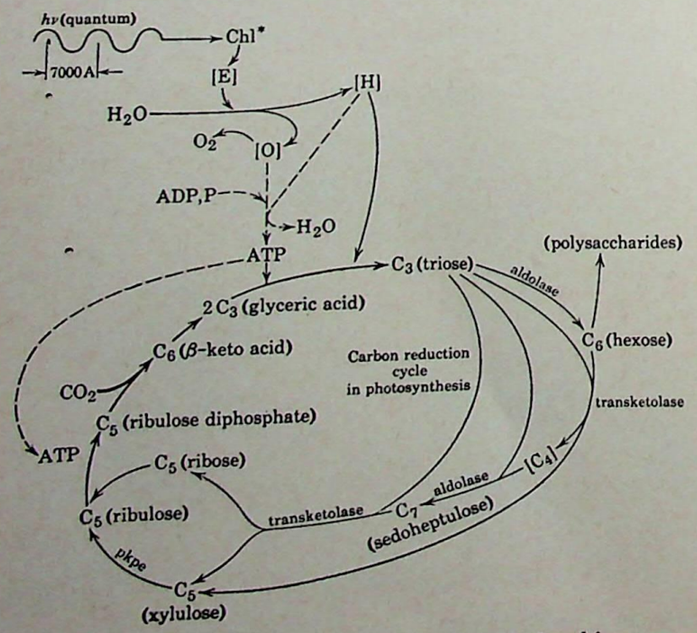


FIG. 16. Suggested cyclic scheme for relationships in photosynthesis.

Clearly the ATP and the TPNH (energy sources for the photosynthetic cycle) must come ultimately from the light. The remainder of this paper is devoted to this problem.

How does the light, which is absorbed by the chlorophyll, produce these two substances (ATP and TPNH) which are known to be required for carbon reduction? Before going into the details of a possible answer, consider some pictures of the apparatus which does it. Figure 17 shows a photomicrograph of liverwort tissue; one can see the cells, the cell walls, and the chloroplasts in which the chlorophyll is distributed very nicely inside the cells. Figure 18 shows isolated chloroplasts from spinach. (These are bigger chloroplasts and they have been isolated in sucrose solution.) All of this carbon reduction, oxygen evolution, phosphate production can be done with these chloroplasts removed from their natural habitat inside a cell. However, in order for that to be possible at anything approaching the rates at which it occurs in the living cell, one must add cofactors, some of which are heat stable, some of which are heat labile, and some of which are unknown but which are obtained out of the sap of the cells. In any case, this whole process can be done outside the cell.

Figure 19, from the work of Steinman and Sjöstrand,[3] shows an electron micrograph of a chloroplast. The picture on the right is shown at a higher magnification. The outstanding features of the chloroplast structures are lamellae, discussed in the following.

For some twenty years, it has been possible to carry out the photochemical evolution of oxygen by isolated chloroplasts using a suitable hydrogen acceptor such as ferrocyanide or quinone. This is called the Hill reaction. In the last five years, by preparing the chloroplasts in a manner which presumably does not destroy a chloroplast membrane or perhaps precipitates enzymes from the cytoplasm onto the chloroplast (i.e., preparing the chloroplasts in salt or sugar solutions), our laboratory has been able to carry out two other reactions with the chloroplasts. These reactions are the reduction of CO_2

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

as well as the evolution of oxygen, and, finally, the production of ATP by illumination of the chloroplasts. These three reactions, carbon-dioxide reduction (or, one step further back, the production of reduced pyridine nucleotide rather than CO_2 reduction), ATP production, and oxygen production are the three processes that one now can accomplish with the chloroplasts. The reduction of CO_2 requires two of the items and the evolution of oxygen may or may not require ATP.

How many of these things can be done simultaneously by the chloroplasts? In a recent conference,[4] it became evident that all of the pair combinations of these processes (i.e., CO_2 reduction, ATP production, and oxygen production) could be demonstrated.

It has been demonstrated that one could make one mole of pyridine nucleotide for every atom of oxygen produced. Simultaneously, one can demonstrate the production of one mole of ATP for every equivalent of reduced pyridine nucleotide produced. (One can demonstrate now that one mole of ATP is created for every equivalent of oxygen produced simultaneously.) This is something beyond the oxidative phosphorylation about which Lehninger writes (p. 136); that is, the oxidative phosphorylation would be the production of ATP by a recombination of TPNH and intermediate oxidant. It now appears that all three of these things can be reproduced equivalently at the same time.

The apparatus which does this in the plant has been shown in three magnifications—the whole chloroplasts in the cells, the chloroplasts outside the cells, and, finally, the lamellar structure of the chloroplasts as seen by electron microscopy. Studies of this lamellar structure have resulted in a particular conclusion which is sufficiently general to be stated; namely, the chloroplast lamellae seem to be (no matter what plant cell is investigated) disk-like in character; they seem to be connected at the edges to form a hollow disk—this is the lamella. The lamellae are quite long, about 2000 A

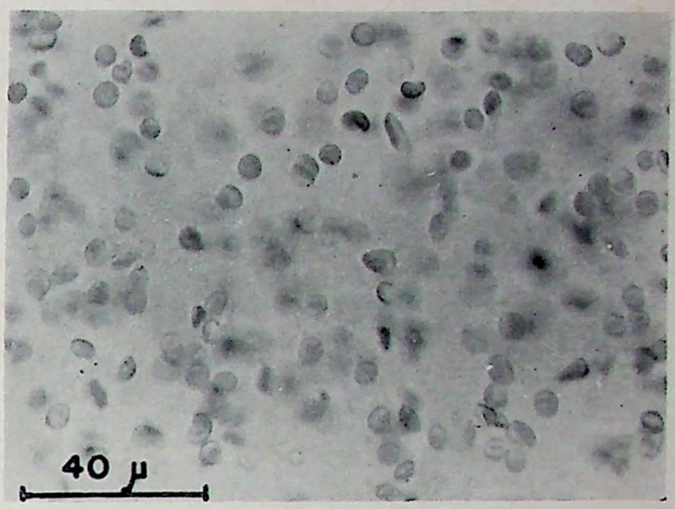


FIG. 18. Photomicrograph of spinach chloroplasts.

in spinach chloroplasts. The lamellae do not appear in the chloroplast in the absence of chlorophyll or protochlorophyll. If, in some way, one prevents either the formation of protochlorophyll or of chlorophyll, one prevents the appearance of well-developed lamellae. Protochlorophyll alone will induce in cells which are normally capable of making them structures which look like these lamellae.

The possible function of this lamellar structure of the chloroplasts is discussed now. The basic problem of photosynthesis can be reduced to the problem of converting a 35- to 40-kcal quantum of energy into some chemical potential. In order to do this, one presumably has to find a reaction, which will take up 35 kcal at one time. The products of this reaction must not back-react. The 35 kcal are a great driving force for the back-reaction, so there must be some mechanism provided in the apparatus to prevent it.

There are a number of other difficult requirements which must be fulfilled in this quantum-conversion process with respect to the time constants involved. For

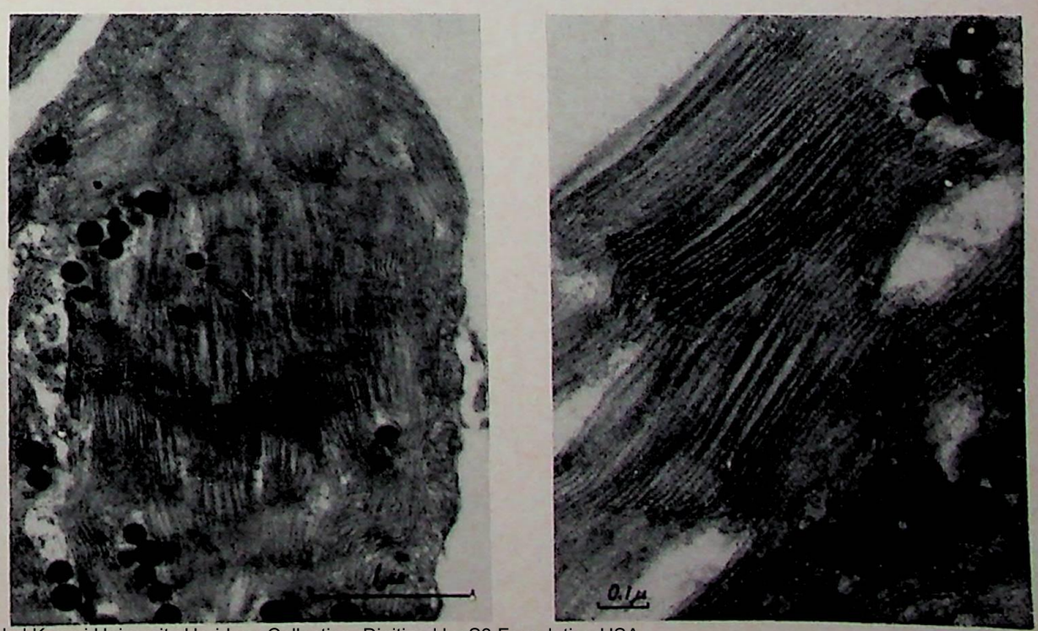


FIG. 19. Ultrastructure of chloroplasts [from E. Steinman and F. S. Sjöstrand, Exptl. Cell Research 8, 15 (1955)].

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

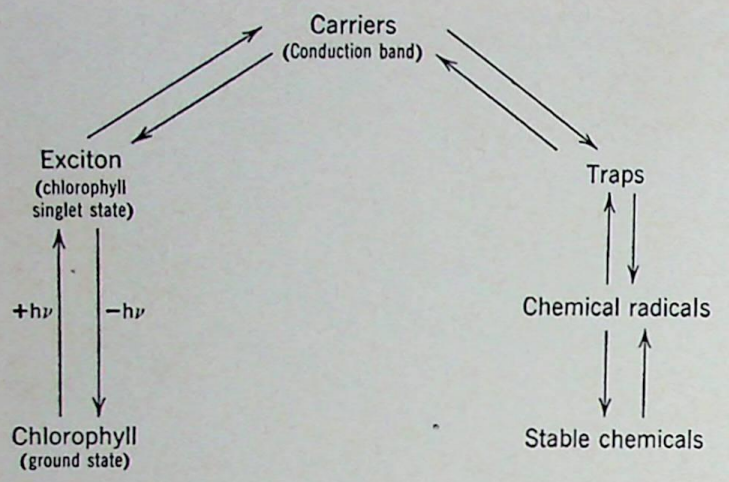


FIG. 20. Hypothetical scheme for light-energy utilization on chloroplasts.

example, following the absorption of the quantum, there must be a very efficient way in which the excited state of chlorophyll can be converted very quickly into a long-lived chemical potential because of the efficiency of the over-all process, regardless of whether one believes the maximum efficiency to be 30 or 60%.

There are a number of approaches to this problem which are based upon ordinary statistical solution photochemistry. In the past, I have looked for reactions unique to chlorophyll that might conceivably be used to store this 35 kcal of energy, such as reduction of chlorophyll to bacteriochlorophyll (that is, adding two more hydrogens to the chlorophyll molecule). Perhaps the process can be accomplished in the reverse manner, taking off two hydrogens from the chlorophyll molecule, making protochlorophyll, and hanging the hydrogen atoms onto something else. These are possible reactions of chlorophyll.

As a matter of fact, our laboratory demonstrated both of these reactions some years ago. More recently, and more elegantly, they have been demonstrated by Krasnovskii[5] in systems that are more nearly related to those which one finds in the living organism.

As a result of a variety of requirements, as a result of the recognition of this highly organized apparatus in which chlorophyll occurs in the chloroplast, and as a result of the failure to solve the problem with solution photochemistry, our group has turned to the notions of cooperative phenomena of organized systems such as those which are represented by barrier-layer cells in physics. Our group is trying to visualize how a lamellar structure such as this conceivably might be an unsymmetrical layer in which one could generate, by the absorption of light, an oxidant and a reductant, on opposite sides of the layer, so that they could not back-react easily and could persist for a long period of time. These substances (reductant and oxidant) should live long enough, by chemical standards, to be taken up efficiently, on the one hand, by electron acceptors to go on to make the reduced pyridine nucleotide, and, on the other hand, by electron donors to make molecular oxygen. I should like to present a proposal which fulfills all of the necessary requirements of the molecular interactions together with the need for conductivity (electrical conductivity) in certain parts of the lamellae and the consequent separation of charges.

The basic proposal is given in Fig. 20 which suggests how these lamellae achieve this energy conversion. Chlorophyll in the ground state absorbs light which brings it to its lowest singlet excited state. The excited state can move around among the chlorophyll molecules by resonance transfer (exciton migration) until a point is reached where ionization occurs. Then charge separation can take place. The exciton can be visualized as a charge-pair which cannot move separately—a positive charge and a negative charge which must move together. When a suitable point in the chlorophyll lattice is reached where the charges can be uncoupled so that they can move separately, there is a conduction band. The electrons can move in one direction and the holes, or positive charges, in another. The electrons and holes move around until they find suitable places of lower potential energy into which they fall, and there sit for times sufficiently long so that suitable chemicals can come up and take off electrons, on the one hand, and the positive holes, on the other. This leads to chemical reactions which then produce stable chemicals such as a pyridine nucleotide and perhaps hydrogen peroxide, or something else of that sort, ultimately going on to the final products.

With this concept, consider how the structure of the lamella may be interpreted in terms of the molecular constitution. It is suggested that this layer is made up of at least four components [Fig. 21(b)]. The protein enzymes involved in carbon-dioxide reduction are on the outside of the disk. The protein enzymes on the inside of the disk are involved in oxygen evolution. The separation of the two processes (carbon-dioxide reduction and oxygen evolution) is achieved by a layer of chlorophyll packed in the characteristic aromatic way. This is a very characteristic pattern of packing. The aromatic rings do not pile flat on themselves; they lie at an angle, approximately 45° to the stacking axis. This type of packing is suggested for chlorophyll.

Figure 21(a) represents chlorophyll molecules tipped this way. Packed between them are carotenoids and the phospholipids. The proposal is that, after absorption, the exciton can migrate around among a few of these chlorophyll molecules to find a suitable point of ionization where the electrons may move in one direction and the positive holes in the other. Thus, one side leads to oxygen production, and the other to the reduction of carbon.

What kind of experimental evidence might detect this kind of system? Electrodes cannot be placed on these lamellae; they are too small. But one part of this scheme is susceptible to experimental observation—namely, the trapped holes and trapped electrons. It is expected that

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

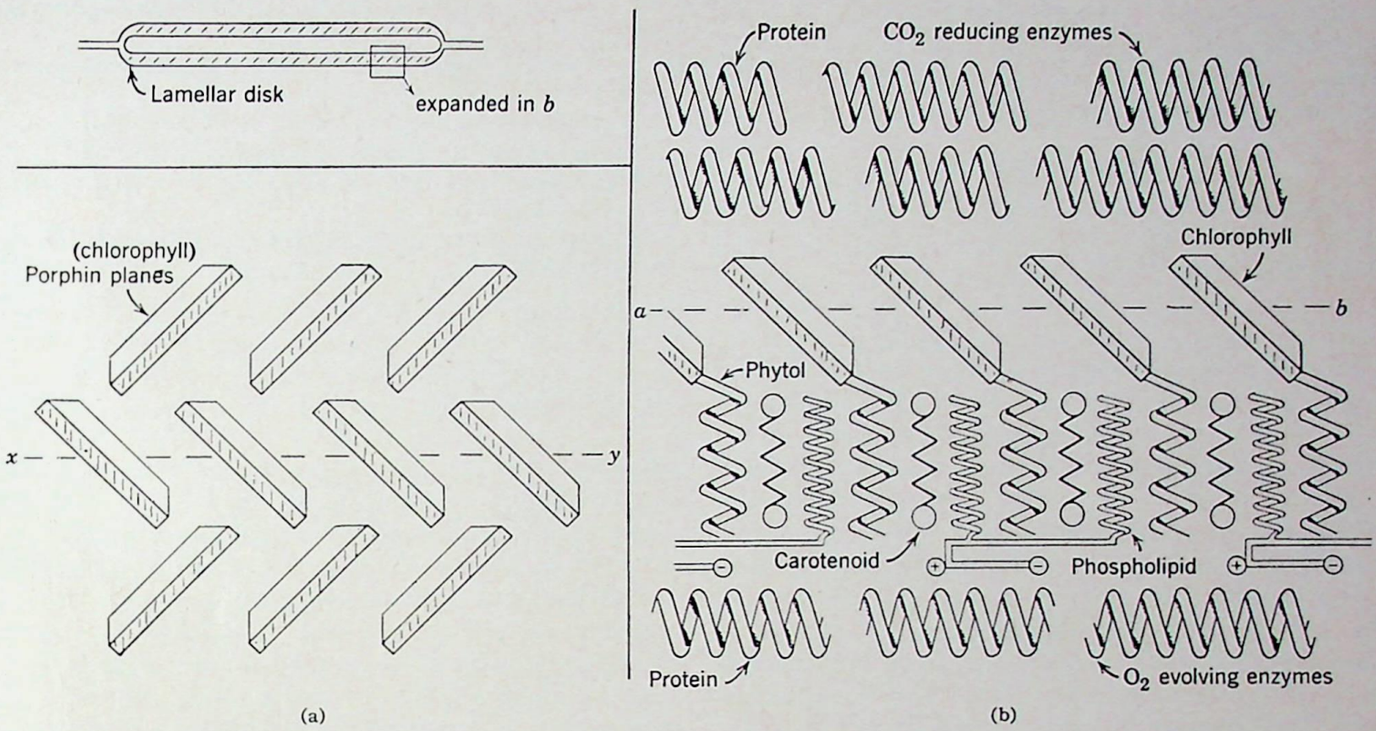


FIG. 21. Schematic representation of possible molecular structure for a lamella.

these trapped electrons are single, trapped electrons and, therefore, are detectable by paramagnetism. Our laboratory set out to search for photoinduced paramagnetism in the chloroplasts. Figure 22 shows the results of that search.[6–8] This is an illustration of electron spin-resonance signals for illuminated whole spinach chloroplasts at 25°C and at −150°C. (Similar signals, at least at room temperature, were reported from the St. Louis laboratory by Townsend.) The fact that one can get the signals at −150°C, either in chloroplasts or in algae, indicates that their production is not an enzymatic process.

The next question that may be asked is: "How fast can the signals be produced at −150°C as compared to 25°C? This is shown in Fig. 23. At 25°C, when the lights are turned on, the signals grow more rapidly than the instrument can follow them. At −150°C, the signals grow equally rapidly. The difference lies in the rate of decay of the signals. They have a complex decay—partly rapid and partly slow. At room temperature, the decay is rather rapid. At −150°C, there may be a rapid decay, but most of it is slow. This at least eliminates the possibility that the signals result from enzymatic formation. The questions remain, could the signal result from a triplet state—that is, a paramagnetic excited chlorophyll—or could the signal be the result of a

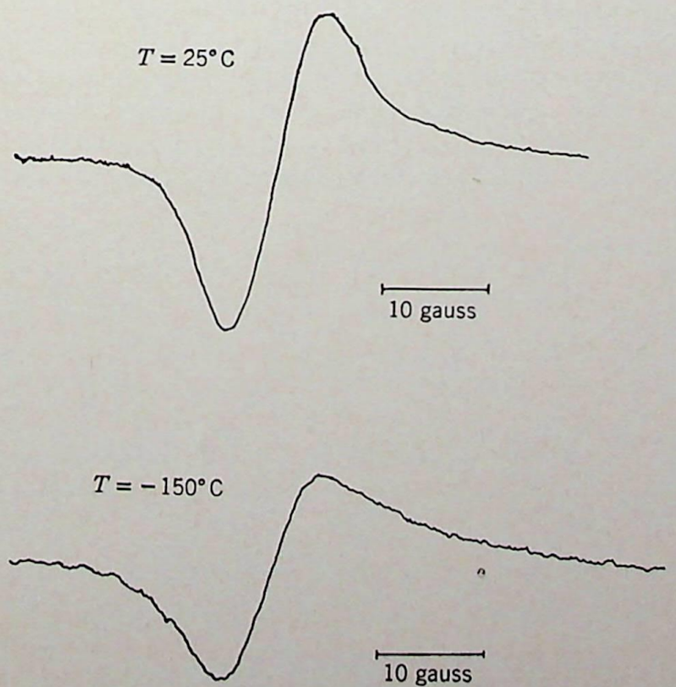


FIG. 22. Light signals from whole spinach chloroplasts.

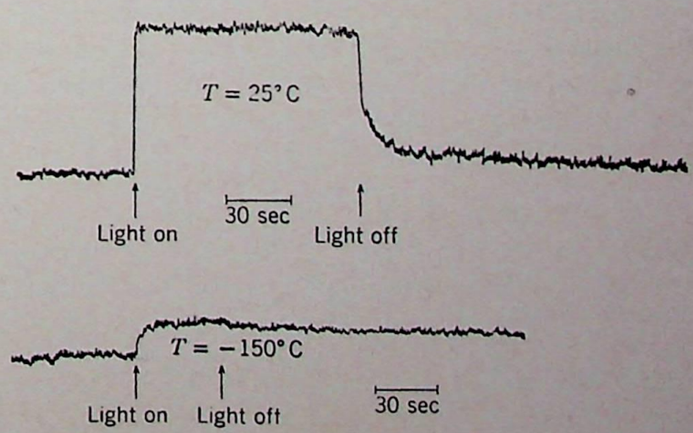


FIG. 23. Signal growth and decay time curves of whole spinach chloroplasts at 25°C and at −150°C.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

photodissociation of chlorophyll, or something very closely associated with chlorophyll, to form chemical radicals, which process can take place at −150°C? These questions are discussed in the next article.

BIBLIOGRAPHY

[1] V. Moses and M. Calvin, Proc. Natl. Acad. Sci. U. S. 44, 260 (1958).

[2] E. Racker, Nature **175**, 249 (1955).

[3] E. Steinman and F. S. Sjöstrand, Exptl. Cell Research **8**, 15 (1955).

[4] Brookhaven National Laboratory Biology Conference No. 11 (June, 1958).

[5] A. A. Krasnovskii, J. chim. phys. (to be published).

[6] M. Calvin and B. P. Sogo, Science **125**, 499 (1958).

[7] P. B. Sogo, N. G. Pon, and M. Calvin, Proc. Natl. Acad. Sci. U. S. **43**, 387 (1957).

[8] P. B. Sogo, M. R. Jost, and M. Calvin, Radiation Research (to be published).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

18
Free Radicals in Photosynthetic Systems*

MELVIN CALVIN

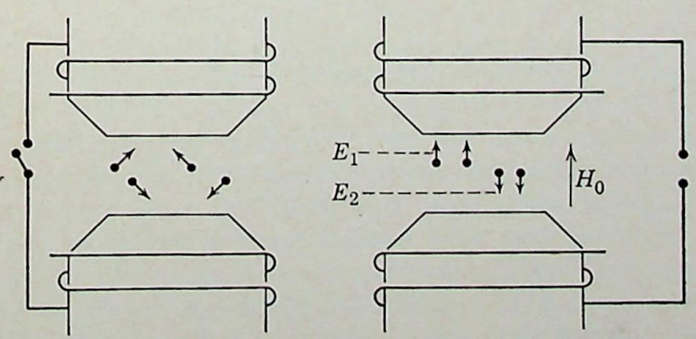
Department of Chemistry, University of California, Berkeley 4, California

ONE bit of evidence introduced in the preceding paper concerned the possibility that the quantum-conversion act of photosynthesis might be a production of unpaired electrons which were trapped successively and then handed down to other acceptors (chemical acceptors) to do their job by ultimately reducing carbon dioxide.

That particular evidence was not explained. It simply was stated that there was evidence for unpaired electrons. The nature of this evidence now may be explained. These curves previously shown are known as electron spin-resonance absorption spectra. This article describes electron spin resonance, how it may be used for the detection of unpaired electrons in biological systems, how it has been used, and how it further might be used to identify the nature of the unpaired electrons that might and do occur in biological systems.

Electron spin resonance is another spectroscopic method. Zavoisky[1] was the first to make use of it to detect unpaired electrons in physical systems. It is based on the principle that an electron has a spin giving it a magnetic moment such that, when placed in a magnetic field, the electron orientates itself with respect to the magnetic field in certain specific directions. In the case of the electron, the spin is said to be one-half of a unit, and this leads to only two possible orientations of the electron in an external field—with and against the field.

Figure 1 shows the diagrammatic representation of this situation. On the left is the case in which the electrons are in between the pole faces of a magnet, but the electric current is not flowing yet and the electron magnetic moments are arranged randomly. When the field is turned on (as on the right), some of the electrons orientate themselves with the external field, and some against it. These two orientations do not have the same energy. The energies are designated as E_1 and E_2 in the figure, and the difference between them is equal to the product of the magnetic moment μ_0, the gyromagnetic ratio g_0, and the value of the external magnetic field H_0. This difference in energy can be expressed in terms of the frequency $h\nu$, and for an electron in a field of about 3000 gauss, the wavelength corresponding to this transition is about 3 cm.



$$E_2 - E_1 = \mu_0 \cdot g_0 \cdot H_0 = \Delta E = h \cdot \nu = h \cdot c/\lambda$$

$$\lambda = \frac{h \cdot c}{\mu_0 \cdot g_0 \cdot H_0} \qquad \mu_0 = 0.927 \cdot 10^{-20} \qquad H_0 = 3300 \qquad \lambda \sim 3{,}2 \text{ cm}$$

FIG. 1. Energy states of free electrons in an external magnetic field.

One may observe unpaired electrons by shining them with electromagnetic energy of a wavelength corresponding to the transition, and by watching for the absorption of this characteristic frequency, ν, when the magnetic field, H_0, is varied. One then determines the magnetic field at which absorption occurs.

Figure 2 indicates the manner in which the experiment is done, showing the 3-cm generator and the external magnetic field. In the presence of the magnetic field, some of the electrons are oriented parallel with the field, some antiparallel. There will be more in the lower (parallel) state than in the upper (antiparallel), so there results a net energy absorption as transitions between states occur.

Most of the apparatus in use today does not record direct absorption, recording instead the derivative of this absorption. The two graphs in the preceding paper (Figs. 22 and 23) are these derivatives of absorption rather than the absorption itself.

This experiment thus provides a method for the specific detection of unpaired electrons (since paired electrons mutually each quench the magnetic moment of the other). It provides a highly sensitive device, more sensitive in general than that provided by any mass susceptibility measurement, because the latter always must correct for the diamagnetic material in which the electron is buried. In recent years, this device has been used more and more in chemistry and now is being used on biological materials to detect the presence of free radicals and to determine whether or not free radicals

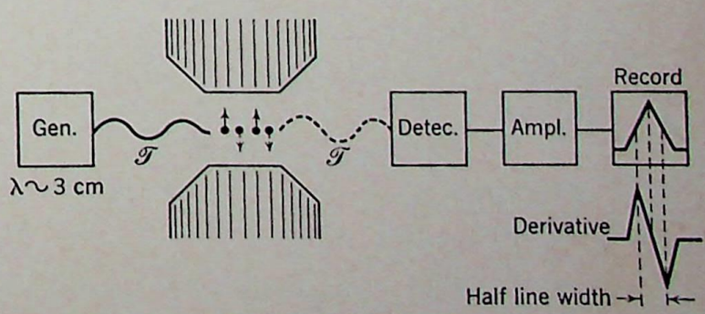


FIG. 2. Absorption of 3-cm waves resulting from transition of free electrons between energy states in an external magnetic field.

* The work described herein was sponsored by the U. S. Atomic Energy Commission.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

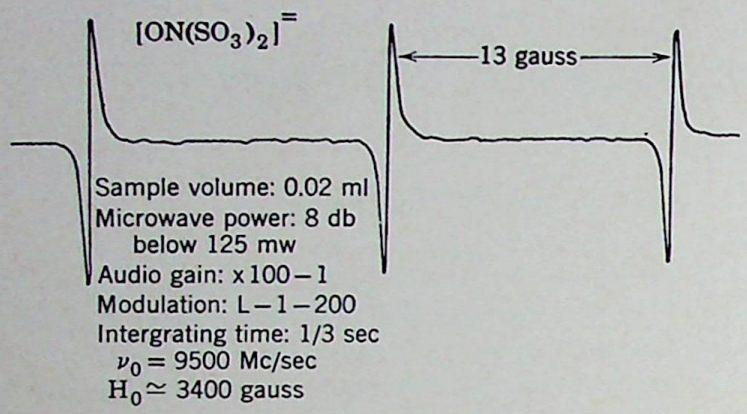


FIG. 3. ESR spectra of nitrosyl disulfonate [from Varian Associates Technical Information Bulletin, No. 2, 11 (1958)].

and unpaired electrons are participants in biochemical transformations.

The measurements obtained by this method not only can reveal how many electrons there are, but also, by the nature of the absorption record, may reveal something of the environment in which the electrons are situated. One recalls that the energy of the transition is dependent upon the field that the electron sees—that is, the external field due to the magnet, plus any magnetic field which the molecule may cause. The molecule itself is made up of nuclei and other electrons. Most of the other electrons, of course, are paired off, but the nuclei may have magnetic moments. Some have magnetic moments that produce magnetic fields with which the unpaired electrons may interact. If the odd electron sees not only the external magnetic field, but also that of the molecule, a spectrum similar to that shown in Fig. 3 from Varian Associates[2] may be observed. In this case, the molecule is one which has in it a nitrogen atom, which has a nuclear spin of one unit. This means that the nitrogen nucleus can take up one of three orientations in the external field—parallel, antiparallel, or normal. The electron near the nitrogen atom actually may be subjected to three different magnetic fields. In one configuration, the magnetic field of the nucleus is added to the external field. In another, it is subtracted from the external field. In the third case, there is no effect on the external field. So instead of one there should be three peaks. Figure 3 shows the three peaks attributable to the interaction with the spin of the nitrogen nucleus: one in the middle, and one on each side resulting from the three possible orientations of the nitrogen nucleus.

Nitrogen is not the only nucleus having a magnetic moment. One of the most common in organic substances is hydrogen. Hydrogen has a spin of a half, and it also

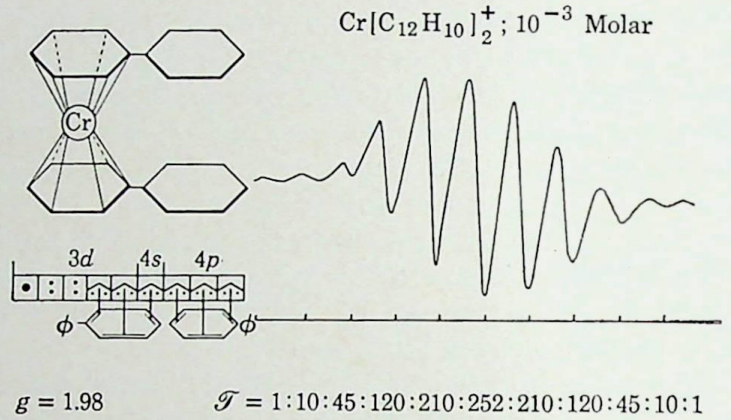


FIG. 5. Paramagnetic-resonance absorption spectrum of bisdibenzene chromium cation.

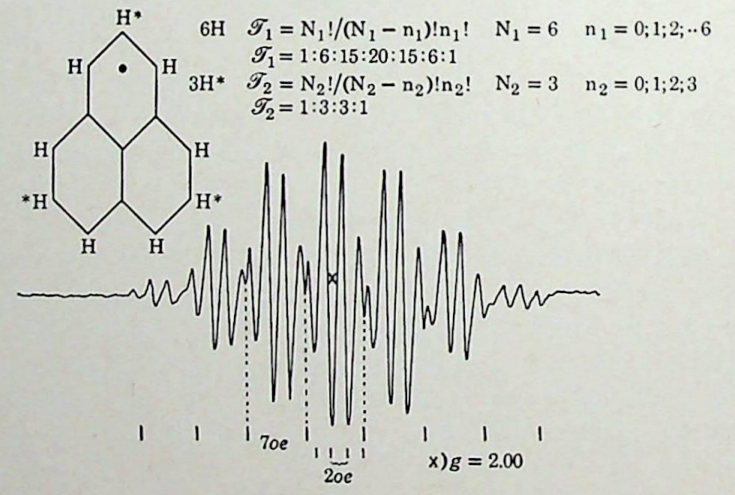


FIG. 6. Paramagnetic-resonance absorption spectrum of perinaphthenyl radical.

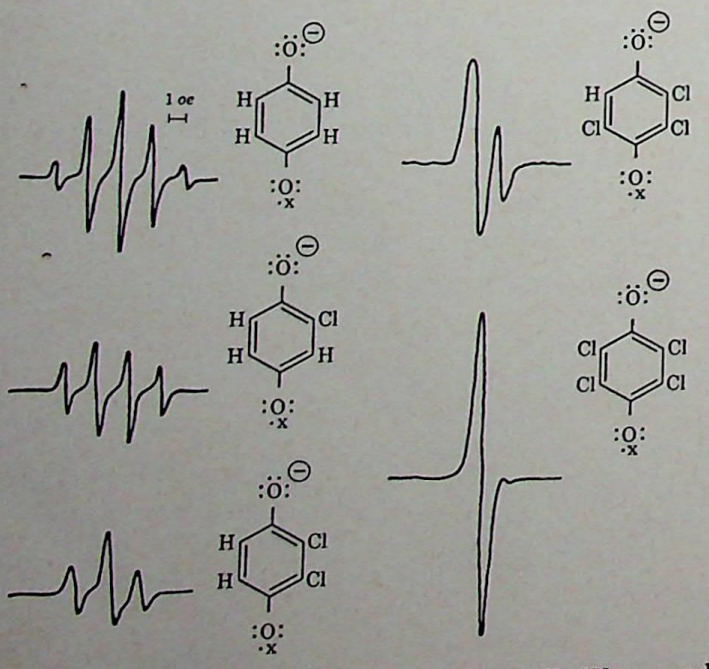


FIG. 4. ESR spectra of chloroquinones [from J. E. Wertz and J. Vivo, J. Chem. Phys. 23, 2441 (1955)].

can influence the spin spectrum of an odd electron in the case where molecules are constructed such that the electron interacts with these protons.

Figure 4, from work by Wertz and Vivo,[3] shows the case for tetrachlorohydroquinone (lower right). The resonance field is not modified by any of the atoms in the molecule since their nuclei are all of spin zero (no magnetic moment), and thus only one line obtains. If, however, one removes one of the chlorines and replaces it by a proton, and if the electron can interact with that proton, the proton, having two possible orientations in the magnetic field, splits the resonance into two lines.

How many different arrangements can one obtain to

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

change the external magnetic field if two protons are present? Both of the magnetic moments may be had with or against the field, or they can be had one with and one against the field. There are, therefore, three different ways in which these two protons can be arranged with respect to the external magnetic field. The third situation is twice as probable as either one of the other two. Therefore, the middle peak should be twice as high as either of the two outside ones. Indeed, this is the case. In both cases, they produce no net modification of the external field. If three protons are present, there are four possible arrangements, and one gets four peaks; with four protons, there are five possible arrangements, and the amplitude ratios are the ordinary binomial coefficients, as seen in the figure.

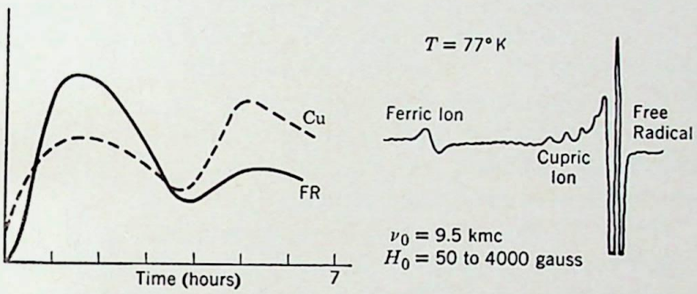


FIG. 7. ESR spectra from coenzyme-*A* dehydrogenase plus substrate [from H. Beinert, EPR Talk No. 10, Varian Associates (1958)].

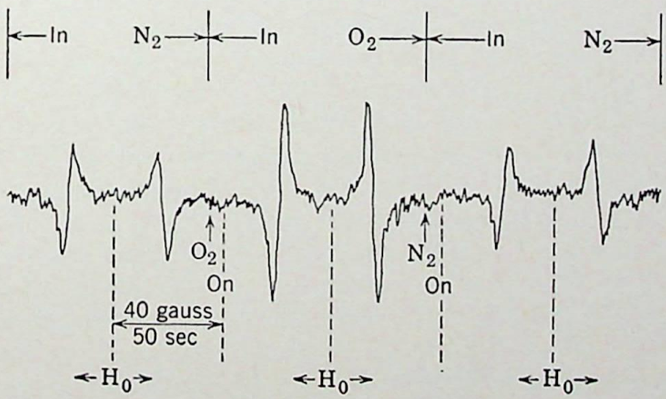


FIG. 8. Dark signal from *Rhodospirillum* methanol extract at room temperature

Not only may one determine, therefore, the number of odd electrons present in a system, but also one may ascertain something about the kind of an environment in which they are. Figure 5 shows the compound, dibenzene-chromium cation, which has one unpaired electron. The question of interest is the location of this unpaired electron. It was found, in our laboratory, that this unpaired electron of the chromium actually can see the ten protons of the two benzene rings of this compound associated with the chromium atom. Ten protons have eleven possible arrangements, and there should be eleven peaks in the curve. Those on the outside are very weak, because the probability of having all ten protons oriented in the same direction is very small as compared with the mixed configurations. One can determine, therefore, something about the environment of the transition elements by this method.

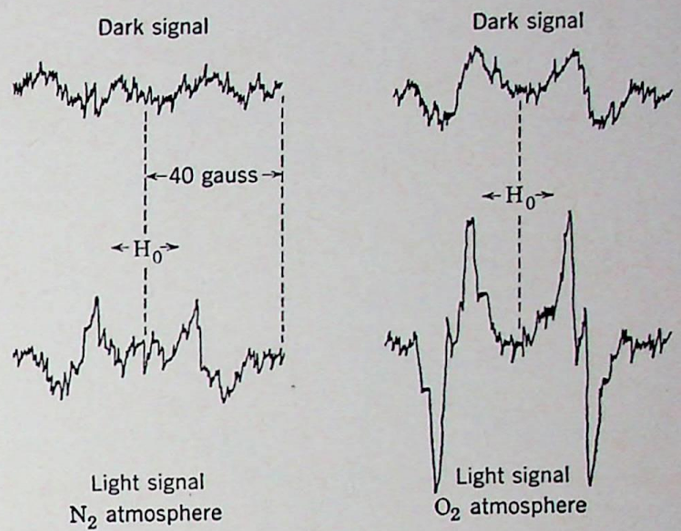


FIG. 9. Light and dark signals from *Chlorella* methanol extracts in two atmospheres at room temperature.

As a final chemical example, consider a perinaphthalene radical—a beautifully symmetrical radical (Fig. 6). During the preparation of the perinaphthalene, it accidentally became oxidized, and a very formidable absorption spectrum was obtained. The odd electron sees two different kinds of protons. On careful examination, one sees that there are seven groups of lines—each group a quadruplet. This means that there are six protons of one kind and three of another. The electron interacts more strongly with the six-proton group than it does with the three-proton group.

How much of this can be used for the investigation of biological free radicals? Beinert[4] has been working with a fatty acyl CoA which he found produced a transient, colored intermediate when mixed with its substrate. This, he proposed, was a free-radical intermediate, and when it was examined at Stanford University and at Varian Associates, an electron spin-

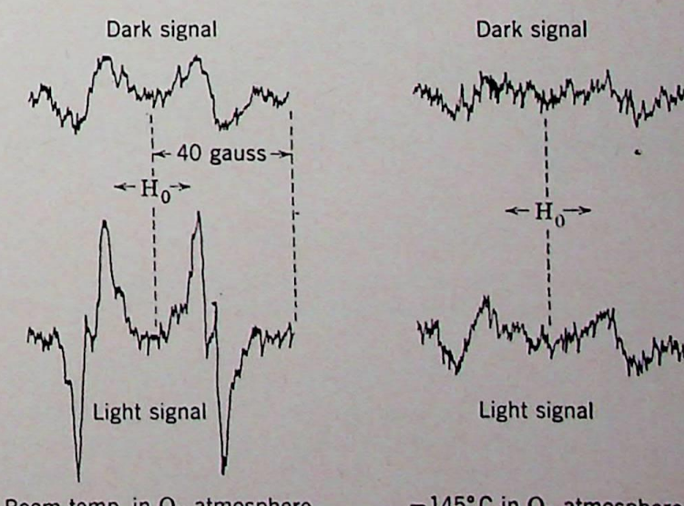


FIG. 10. Light and dark signals from *Chlorella* methanol extract at two temperatures.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

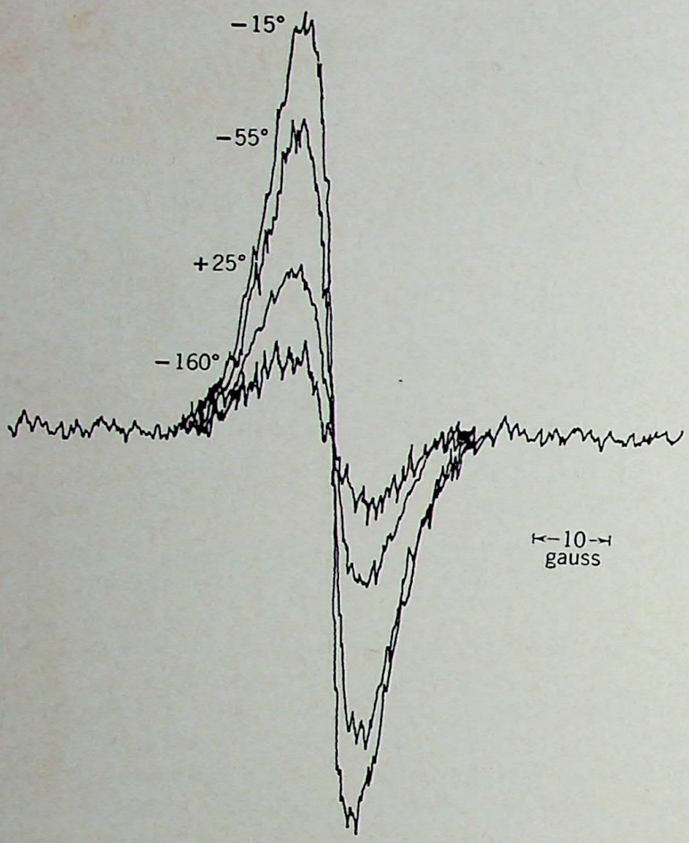


FIG. 11. ESR signals from *Rhodospirillum rubrum*, 5 min continuous illumination.

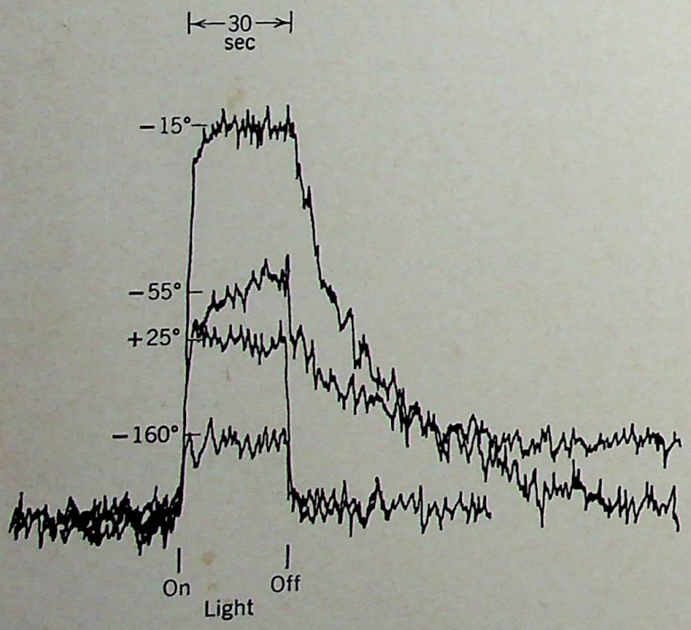


FIG. 12. Rise and decay of ESR signals from *Rhodospirillum rubrum*.

resonance signal was seen which could be distinguished from the resonance owing to the Cu^{2+} in the enzyme (Fig. 7). Commoner[5] and co-workers presented evidence of free-radical intermediates in the reactions of alcohol dehydrogenase and of cytochrome oxidase when these enzymes are brought into contact with their substrates.

How is this kind of a signal of a free radical distinguished from that shown in the preceding paper (Fig. 22)? Firstly, radicals of this kind do not fade, at low temperatures. They are frozen in. That is exactly what happens upon cooling the system of chloroplasts as described earlier. The fact is that structured systems can be found wherein the signal can be induced at a low temperature, but wherein it also fades at a low temperature at a very high rate.

The occurrence of a signal owing to an oxidation mechanism alone is illustrated by Fig. 8, which shows the behavior of a methanolic extract of *Rhodospirillum* when exposed to oxygen and nitrogen, alternately. When a similar methanolic extract of *Chlorella* is illuminated first in an oxygen and then in a nitrogen atmosphere, the increased signal of the former (Fig. 9) is considered the sum of the contributions of an oxidative, a photooxidative, and an odd electron produced and trapped in a free radical. The relative contributions of these processes are apparent from the figure. Figure 10 shows how cooling, even in oxygen, reduces the signal amplitudes, and one sees that cooling to −145°C nearly has eliminated the dark spectrum.

Figure 11 shows the spin-resonance signal for *Rhodospirillum* after 5 min of illumination at the indicated temperatures. At 25°C, it is the next smallest signal;

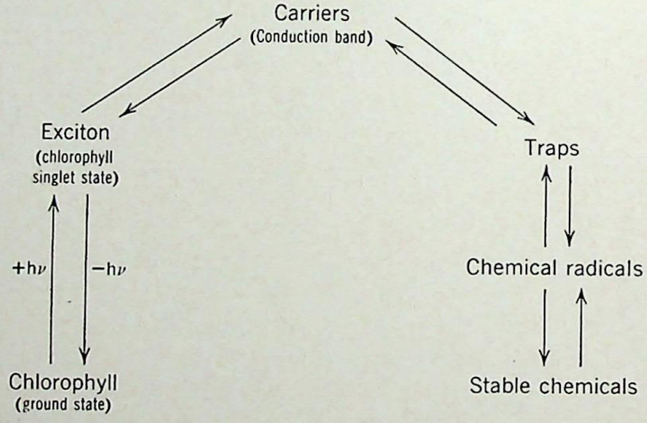


FIG. 13. Hypothetical scheme for light-energy utilization on chloroplasts.

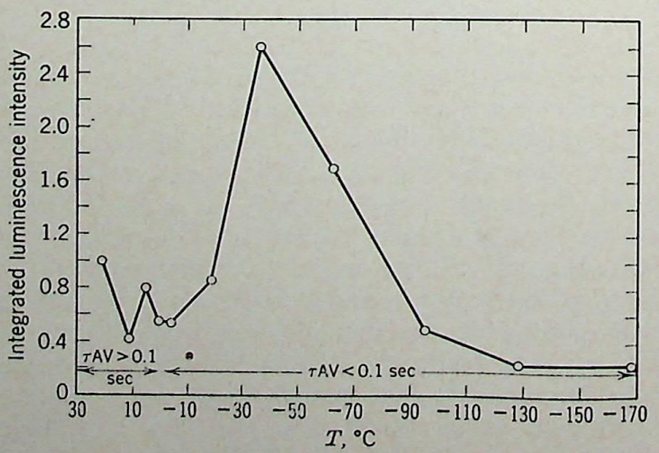


FIG. 14. Integrated intensity of spinach-chloroplast luminescence from 0.0015 to 5.0 sec after excitation, as a function of temperature.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

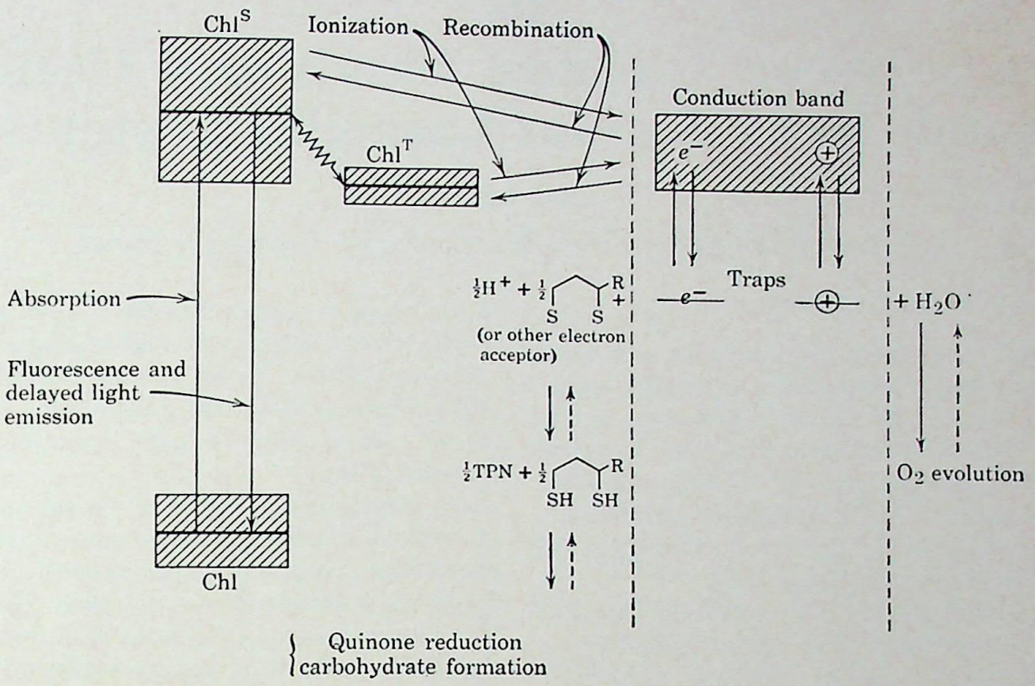


FIG. 15. Proposed scheme for various photochemical processes in photosynthesis.

at −55°C, it is larger; and at −150°C, it is larger yet. When the temperature is further reduced to −160°C, the intensity falls to the smallest shown. The time behavior is shown in Fig. 12. The spectrometer is set on the peak of the signal, and one may observe the rates of rise and of decay of each of these signals at each temperature as a function of time. The room-temperature signal is the next smallest shown, and it rises as rapidly as the instrument can respond. This signal decays immediately after illumination ceases, to within the time constant of the apparatus. There is, however, no rapid decay at −15°C; at −55°C, there is a very rapid rise, a further slow rise, followed by a small component of rapid decay, and a much longer slow decay. At −160°C, all that remains is very rapid rise and very rapid fall. Thus, I believe the possibility has been eliminated that these free radicals are being produced directly by illumination. This behavior must be the result either of untrapped carriers or of trapped carriers in well-shielded traps.

How can this whole sequence of events be accounted for? Figure 13 (discussed in the previous article) indicates the first act, the absorption of light to produce an exciton. The exciton then is converted into conductive carriers, these into chemical radicals, and the chemical radicals lead to stable chemicals. To account for the data shown in Fig. 12, it is necessary to presume that each one of these successive acts has a higher temperature coefficient. The first step has no temperature coefficient, the second may have a very slight one, the third a higher one, and so on. At room temperatures, the energy is transported into the chemical radicals, but these decay very rapidly into stable chemicals by the usual enzymatic process. The possibility that a triplet state exists is eliminated by the nature of the signal. A triplet signal would be a double band very characteristic of two unpaired electrons in the same molecule. A similar sequence of temperature effects is seen in the luminescence intensity curve (Fig. 14) and similarly is accountable, presuming the integrated intensity to be a measure of the back-reaction.

Another representation of the same scheme is shown in Fig. 15. The chlorophyll ground state is represented by a band instead of by a single line—as a result of the interaction of the chlorophyll molecules in these arrays previously shown in electron micrographs—and the chlorophyll excited singlet state actually would be a broader band. The triplet state is shown as a rather narrow band overlapping the excited sequence. The triplet band has not been put directly in line because triplet light emission is not observed.

The foregoing scheme gives a working hypothesis to account for some of the problems, outlined at the onset —that the initial 35 kcal of energy be converted into useful chemical potential.

BIBLIOGRAPHY

[1] E. Zavoisky, J. Phys. U.S.S.R. **9**, 211 (1945).
[2] Varian Associates Technical Information Bulletin, No. 2, 11 (1958).
[3] J. E. Wertz and J. Vivo, J. Chem. Phys. **23**, 2441 (1955).
[4] H. Beinert, EPR Talk No. 10, Varian Associates (1958).
[5] B. Commoner, J. J. Heise, B. B. Lippincott, R. E. Norberg, J. V. Passonneau, and J. Townsend, Science **120**, 57 (1957).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

19

Relation between Exciton Bands and Conduction Bands in Molecular Lamellar Systems*

MICHAEL KASHA

Department of Chemistry, Florida State University, Tallahassee, Florida

1. INTRODUCTION

BIOPHYSICAL investigations on a molecular level turn naturally to the phenomena of energy transfer and electron transfer for the interpretation of the behavior of the ordered structures found in biological systems. The famous Korányi lecture[1,2] of Szent-Györgyi has done much to stimulate interest in these phenomena, although earlier recognition of their importance, e.g., in the photosynthetic system, had been made.

The general tendency in this field at present seems to be the adoption without caution of the ideas and terminology of the solid-state physics of atomic and ionic semiconductors, to the interpretation of biological systems. Since the latter are usually molecular aggregates bound weakly by van der Waals forces (molecular lamellae) or very strongly by overlap forces (intramolecular forces in proteins), it is worthwhile to examine critically the differences between the atomic and molecular array problems, to see how valid is the application of the ideas of solid-state physics to biological mechanisms.

Since most of the treatments of the exciton theory are presented in fairly intricate mathematical terms, an attempt is made here to give a simple but accurate account of excitation-energy transfer as well as of electronic conduction in atomic and molecular aggregates.

2. EXCITONS

The theory of excitons in atomic lattices was developed by Frenkel,[3,4] and elaborated later by other physicists.[5–10] The application of exciton theory to molecular crystals was first made by Davydov.[11,12] A semiclassical treatment of excitons in van der Waals pigment polymers was given by Förster.[13] Qualitatively the same idea was also used by Förster to interpret the spectra of pigment dimers, and has been treated quantum mechanically by Simpson's group.[14] The extension to excitons in pigment aggregates of various geometries has been made by McRae and Kasha.[15] The problem of energy transfer between pairs of randomly oriented molecules was treated quantum mechanically by Förster.[16] All of the foregoing studies make use of the same physical basis, which is described in the following. Several papers of a descriptive nature on the topic of energy transfer have been published,[17–20] mainly from the phenomenological viewpoint.

Three different definitions of the *exciton* have been given, illustrating different aspects of the idea, but not all of which seem equally suitable for a critical understanding:

(1) Exciton states involve excitation of an assembly of atoms (or molecules) in concert, instead of localized excitation of individual species of the assembly.

(2) An exciton can be described by a wave packet traveling through an assembly of atoms (or molecules), and arising from a superposition of exciton states (defined in the foregoing).

(3) An exciton is a neutral excitation "particle," consisting of an electron and a positive hole, traveling together through the lattice.

Each of these definitions is next considered in order to clarify the picture of an exciton.

For a discussion of the first definition of an exciton, consider a simple two-dimensional model of an atomic and a molecular lattice [Figs. 1(a) and 1(b)]. The atomic lattice is assumed to be an ionic one held together by Coulombic forces (for simplicity, the positive and negative ions are not distinguished, the ions are assumed to be isoenergetic in their electronic states throughout, and overlap is assumed to be negligible in the ground state of the lattice). All but one of the ions are shown to be in their ground state, indicated schematically by circles representing spherically symmetrical orbitals.† The ion (2,2),‡ however, is shown with an electron excited to a p-type orbital. This model of the excited lattice forms the basis of what is designated as a *zeroth-order* description, because it is not possible to localize excitation to one atom in such a system. Thus, atom (3,2) might equally well be excited, etc. Abstracting the second row of atoms for simplicity, the following zeroth-order descriptions could be given of singly excited states (where asterisk denotes excitation of atom A_n)

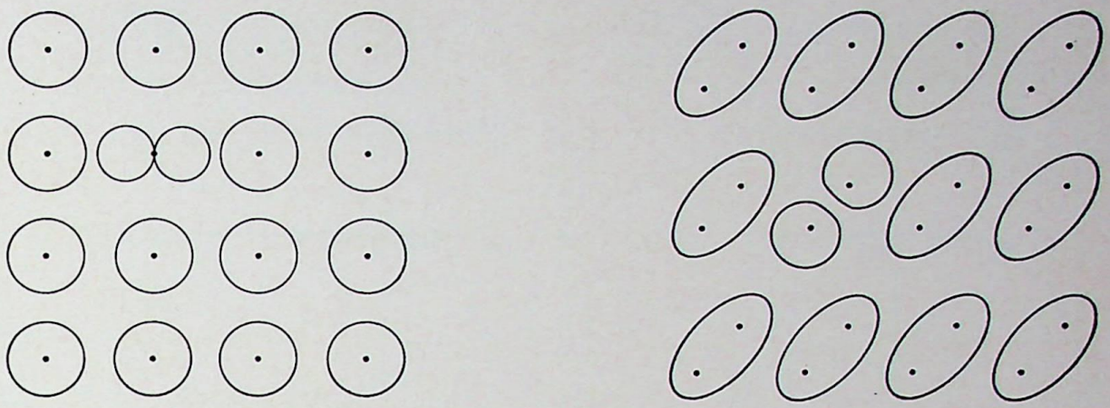
$$\text{(i)} = A_1^* A_2 A_3 A_4,$$
$$\text{(ii)} = A_1 A_2^* A_3 A_4,$$
$$\text{(iii)} = A_1 A_2 A_3^* A_4,$$
$$\text{(iv)} = A_1 A_2 A_3 A_4^*.$$

* This study was made during the Study Program in Biophysical Science held in Boulder, Colorado, in the summer of 1958, under the sponsorship of the National Institutes of Health, Public Health Service, U. S. Department of Health, Education, and Welfare.

† The biologist unfamiliar with the characteristic three-dimension orbital wave functions of atoms and molecules will find it helpful to consult *The Chemical Aspects of Light* by E. J. Bowen [(Clarendon Press, Oxford, England, 1946), second edition], especially pages 68, 100–115, and 139.

‡ Row 2, column 2, from left upper corner.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA



(a) (b)

FIG. 1. (a) Zeroth-order description of a section of an atomic lattice having one atom with an electron in its lowest excited orbital, the remaining atoms with electrons in their ground orbitals. (b) Zeroth-order description of a section of a molecular lattice with one molecule with an electron excited to its next highest orbital, the remaining molecules with electrons in their ground orbitals.

None of these is by itself a suitable description of the excited array, since they could (according to the assumption) all have the same energy, and if they could *interact* with each other, new states which are true stationary states in the quantum-mechanical sense will arise. These new states are the *exciton states* of the first definition; if only the foregoing four atoms were considered, they would have the form

$$\mathrm{I}=\frac{1}{\sqrt{4}}[(\mathrm{i})+(\mathrm{ii})+(\mathrm{iii})+(\mathrm{iv})],$$

$$\mathrm{II}=\frac{1}{\sqrt{4}}[(\mathrm{i})+(\mathrm{ii})-(\mathrm{iii})-(\mathrm{iv})],$$

$$\mathrm{III}=\frac{1}{\sqrt{4}}[(\mathrm{i})-(\mathrm{ii})-(\mathrm{iii})+(\mathrm{iv})],$$

$$\mathrm{IV}=\frac{1}{\sqrt{4}}[(\mathrm{i})-(\mathrm{ii})+(\mathrm{iii})-(\mathrm{iv})],$$

where the signs (see later) result from the requirement (of orthogonality) that the stationary states (I) to (IV) be completely independent. This statement would be true only if it were appropriate to assign an equal weighting to each of the zeroth-order states, i.e., for an infinite array. The present argument is artificial in the sense that it attempts to simplify the picture by abstracting four molecular units from a very large aggregate. It should be noted that each of the stationary exciton states involves all of the zeroth-order descriptions; that is, *each of the stationary exciton states involves excitation of all of the atoms considered as locally excited.* Thus, in the exciton states the excitation is delocalized.

The second definition arises from the possibility, pointed out by Frenkel,[3,4] of construction of wave packets from the stationary exciton states. This is a useful aspect for discussions of electronic energy transfer in atomic and molecular systems. The properties of the wave packet can be calculated by quantum-mechanical methods.[7,9,10]

The third definition can be misleading, so its meaning is examined more fully. In Fig. 1(a), the atomic ion (2, 2) is schematically shown to have an electron in a *p*-type excited orbital. Both the excited atom and the ground-state atom are electrically neutral and electronically symmetrical. The excited atom is produced from a ground-state atom by the polarizing influence of the electric vector of the interacting light wave (the electric vector having been taken to be horizontal, and in plane). While the light wave impinges, a *transitory dipole moment* is induced in the atom, but no permanent charge separation, i.e., ionization, takes place. Thus, the definition of an exciton as an "electron-positive hole" combination can be meaningful largely in a figurative sense, since the exciton involves excited *bound-electron* states of the electron and its originating atom. To actually ionize the "pair" requires the additional energy between the exciton level and the ionization limit.

On the other hand, the transitory dipole induced by the light wave would be directly the cause of the interaction of the zeroth-order states, since all neighboring dipoles could interact with each other in a purely electrostatic way.§ Incidentally, the signs in the exciton states (I) to (IV) should not be taken to mean electron nodes (as in ordinary electronic wave functions), but nodes signifying the phase relation between the transitory dipole moments of the zeroth-order states (see Sec. 4).

The molecular lattice of Fig. 1(b) can be discussed analogously. It is assumed that here van der Waals forces bind the molecules into ordered layers (crystals, or lamellae), and that negligible electron overlap is present in the ground and lowest excited states. The molecular axis is shown inclined at 45° to the layer direction; for simplicity, a two-center molecular orbital is considered. The molecule shown excited (2, 2) has an electronic orbital node perpendicular to the molecular axis. The excited molecule is thus electrically neutral and symmetrical, and is produced from a ground-state molecule by a light wave of the right frequency with

§ In the formal treatment of the theory, the expansion of the intermolecular interaction potential results in a series of terms, of which the dipole-dipole terms are the predominant ones for formally allowed electronic excitation.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

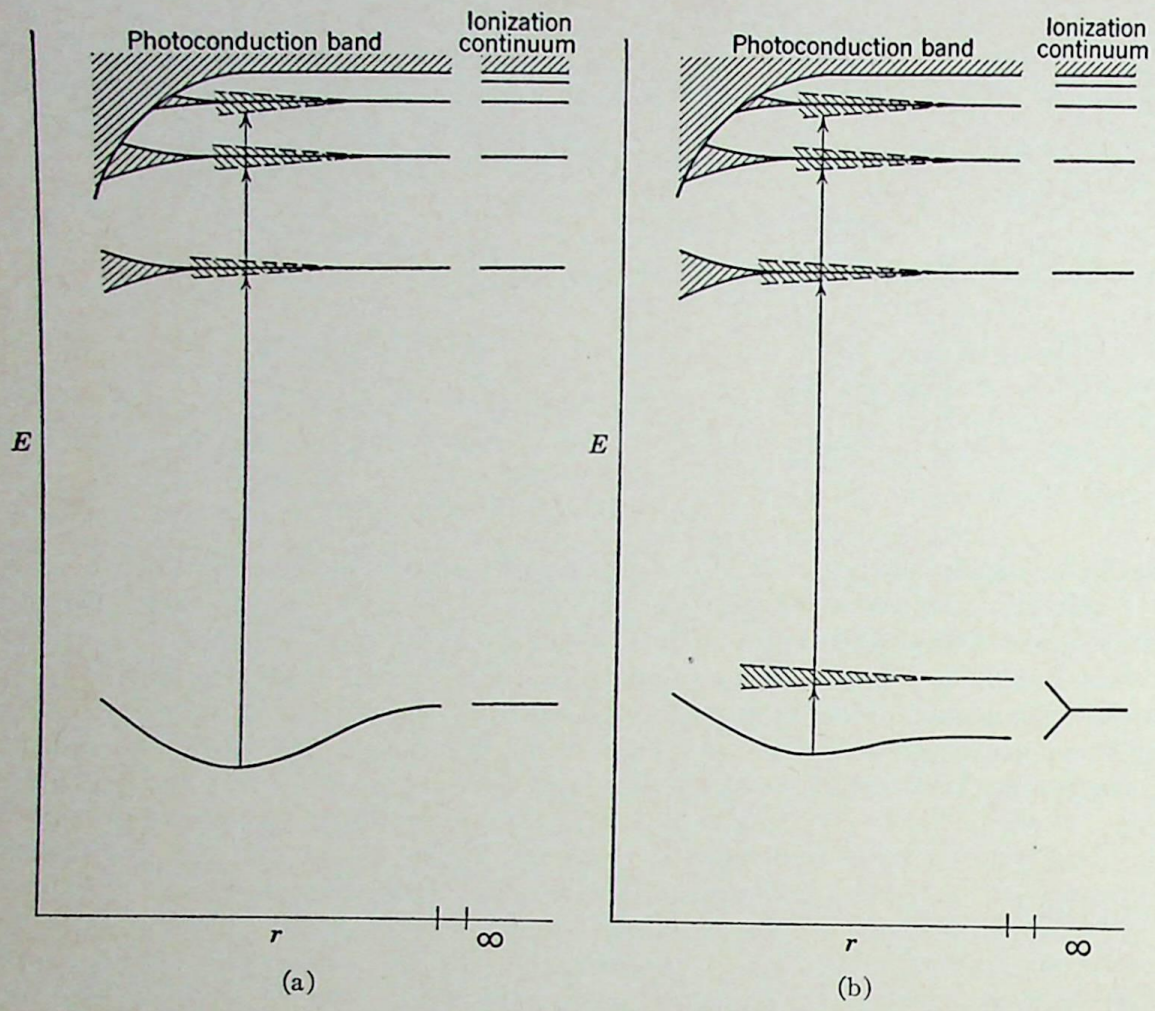


FIG. 2. Schematic energy *vs* interspecies-distance curves for a crystal or aggregate taken as a whole. (a) atomic lattice, (b) molecular lattice or lamellar structure, showing onset of exciton bands and conduction bands. Exciton-band splitting exaggerated, and overlap interaction shown independent of dipole-dipole interaction (cf. Fig. 3).

its electric vector oriented along the molecular axis. The identical symbolic description of the zeroth-order states and exciton states, used for the foregoing atomic lattice model, is applicable to this molecular model.

From the foregoing discussion, it should be clear that exciton states cannot impart electrical conductivity. It is found experimentally that excitation of atomic lattices to exciton states produces no photoconduction of electrons (see Sec. 4).

3. ORIGIN OF EXCITON BANDS AND CONDUCTION BANDS

The origin and relation of exciton bands and conduction bands may be clarified by a discussion of their dependence on interspecies distance. In Fig. 2 are given, schematically, energy *vs* interatomic (interionic) distance (a) and energy *vs* intermolecular distance (b) for the energy levels of a crystal or aggregate taken as a whole.

The atomic array is assumed, as before, to be made up of ions with isoenergetic electronic states, brought together into a Coulombic lattice (positive and negative ions not distinguished). Since the *Coulombic* potential varies as $1/r$, it would be effective at the greatest interatomic separations, and strong ground-state binding results [Fig. 2(a)]. Next in distance dependence would be the long-range $1/r^3$ *dipole-dipole* interaction between excited states. Depending upon the strength of the interaction, each state of the isolated ion would split into an N-fold exciton band (where N equals the number of ions in the assembly). The exciton-band width[15] is proportional to the intensity of the transition in the unit species, and depends upon the orientation of the transition dipole moments (besides the $1/r^3$ dependence, and a rather insensitive function of the number of unit species in the array). Finally, there is the extremely short-range electron *overlap* interaction. This short-range characteristic of overlap forces arises from the exponential fall of electron orbital wave functions in the exterior of atoms (and molecules). In Fig. 2(a), it is assumed that negligible overlap is present in the ground state of the atoms, but that, in the excited state, overlap may occur yielding a conduction band, because of orbital expansion upon high quantum-number excitation. If there is overlap in the ground state, a *valence band* is said to arise, analogous to the conduction band for the excited states.

The point to be noted from the above is that, because of the long-range nature of the dipole-dipole interaction, nonphotoconducting exciton bands can exist discrete from the photoconduction bands. Nevertheless, atomic exciton bands usually lie not far below the photoconduction band of atomic lattices.

The molecular case has some qualitatively different features. However, it is well known that gaseous molecules exhibit converging series absorption in the vacuum ultraviolet (molecular Rydberg series) whose convergence limits represent the ionization potentials for various electrons in the molecule. One could antici-

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

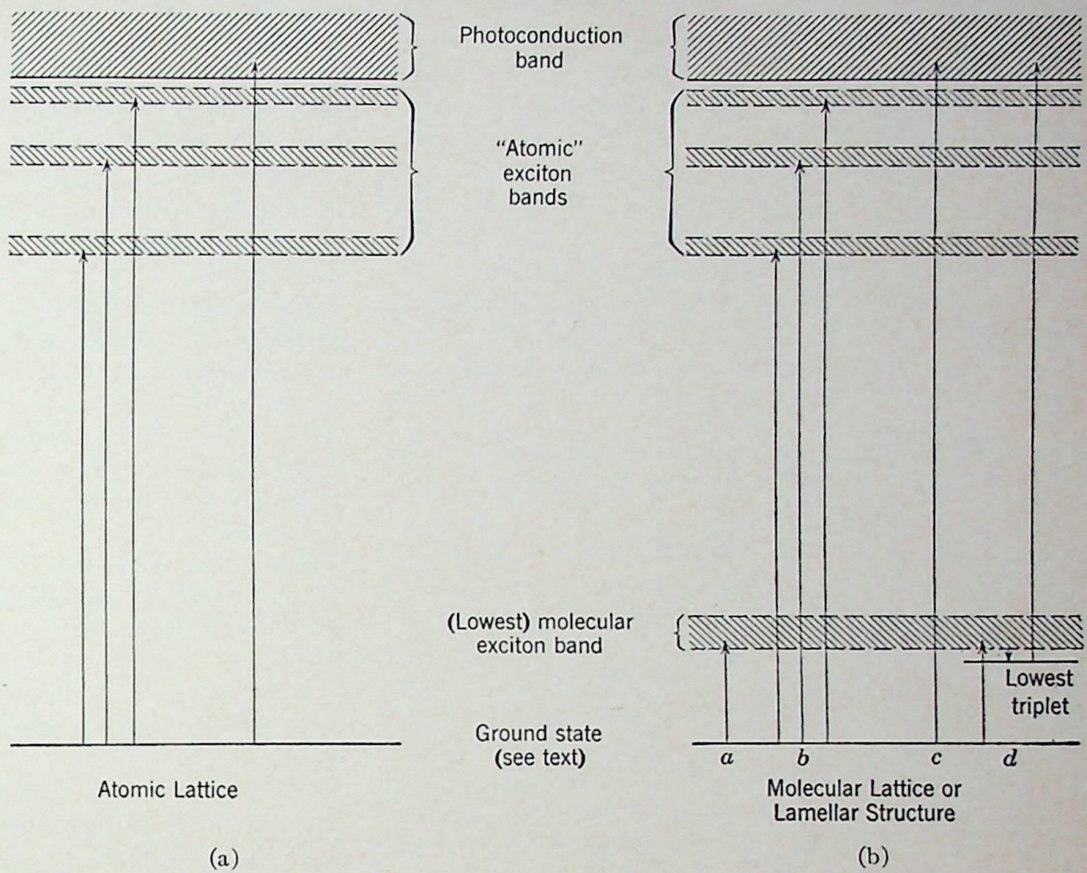


FIG. 3. Energy-band diagram at equilibrium separation for a crystal or aggregate taken as a whole. (a) atomic lattice, (b) molecular lattice or lamellar structure.

pate that dipole-dipole interactions between the electrons in Rydberg levels would give rise to "atomic" exciton bands. These are indicated in the upper part of Fig. 2(b), analogous to the exciton bands of Fig. 2(a).

In addition, a molecule during its formation from atoms [at $r=\infty$ for molecular lattice; lower right corner of Fig. 2(b)] gives rise to excited (unfilled) molecular orbitals. For simplicity, only one filled and one excited orbital are indicated. In the formation of the molecular lattice, therefore, while the ground state becomes slightly lowered in energy through a weak van der Waals interaction, the molecular excited states of the assembly spread by dipole-dipole interaction into an N-fold exciton band (N equals the number of molecules in the assembly). Since molecular transitions are frequently very intense (oscillator strength $f\sim1$), a large band splitting is to be expected (see later). Moreover, since highly conjugated molecules (i.e., pigment or dye molecules) have rather low-energy excited states, exciton bands far from the conduction band should be observed.

4. PROPERTIES OF EXCITON BANDS

The electrically nonconducting characteristic of exciton bands has been referred to above, and the physical basis for this property has been discussed. Some additional characteristics of atomic exciton bands are now mentioned. It is typical of atomic lattices that exciton bands are usually observed just below the conduction band. The optical properties of, e.g., the alkali halide crystals conform roughly to a behavior described by Fig. 2(a), which for the equilibrium interatomic separation yields an energy-level diagram such as Fig. 3(a). The pure alkali halide crystals are properly described as electrical *insulators* rather than as semiconductors, as far as their electronic behavior is concerned. For example, the KCl crystal[8,10] exhibits its first exciton band as a broadened line at 7.6 ev ($\bar{\nu}\sim61\,000$ cm^{-1}, $\lambda\sim1620$ A). Excitation of this band produces no photoconductivity. At 9.44 ev ($\bar{\nu}\sim76\,000$ cm^{-1}, $\lambda\sim1300$ A) photoconduction is observed. Thus, both types of bands are observed in the vacuum ultraviolet not far from each other. No contribution to dark conductivity would be expected at ordinary temperatures from this high-energy conduction band.

Another characteristic of atomic exciton bands is their occurrence in sets which fall into a converging series.[10] This behavior could be anticipated from their atomic character; i.e., the orbitals which would be used in the zeroth-order description of an atomic-lattice exciton are essentially hydrogenic in character, although extending over many lattice points. The series formulas are, of course, modified by the dielectric constant of the lattice, lattice interactions, etc.[10]

The band width for molecular exciton cases has been calculated[15] to be approximately 0.2 ev ($\Delta\bar{\nu}\sim1600$ cm^{-1}) in linear pigment assemblies for an oscillator strength of 1 for the isolated molecule transition, at intermolecular distances of 10 A, and in a suitable geometry. Experimental studies indicate that this is the order of

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

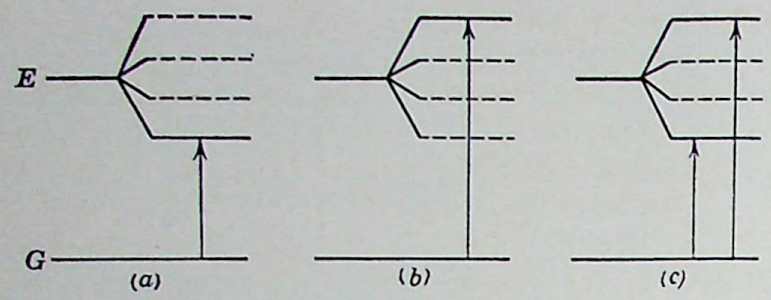


FIG. 4. Selection rules for exciton levels in linear arrays (cf. reference 15), for a hypothetical 4-unit molecular array. (*a*) head-to-tail, (*b*) card-pack, (*c*) alternate-translational-arrays of transition dipoles. Dashed line: forbidden exciton level; solid line: allowed exciton level. Diagram shows dipole-dipole splitting of zeroth-order excited states.

magnitude observed. For weaker optical transitions, or for greater intermolecular separations, smaller exciton-band widths can be expected.

The foregoing comments on exciton-band width are not meant to imply that optical transitions of isolated molecules are necessarily to be broadened in the lamellar aggregates. On the contrary, one of the characteristics[3,4,11,15] of exciton bands is the restriction of optical transitions to only certain components of an exciton band. For example, once more consider the abstracted linear array of four atoms of Fig. 1(a). The exciton states (I) to (IV) can be depicted by vector models representing the dipole moments induced by the light wave with electric vector in the direction of the array,

$$\mathrm{I} \rightarrow\rightarrow\rightarrow\rightarrow,$$
$$\mathrm{II} \rightarrow\rightarrow\leftarrow\leftarrow,$$
$$\mathrm{III} \rightarrow\leftarrow\leftarrow\rightarrow,$$
$$\mathrm{IV} \rightarrow\leftarrow\rightarrow\leftarrow.$$

The dipole-dipole interaction would give (I) the lowest energy and (IV) the highest. Also, the resultant transition moment (vector sum of the individual moments) is finite in the case of (I), but zero for the remaining three states. Thus, the exciton band for this case would be observed to have the structure shown in Fig. 4(a), with only the lowest exciton state permitted in optical absorption. Thus, for this example, a long wavelength or red shift would be observed in the linear aggregate compared with the isolated atomic species, rather than a band broadening. Cases 4(b) and 4(c) correspond to other geometries for a linear array, the former for a card-pack orientation of dipoles, and the latter for an alternate translational orientation. The staggered arrangement of naphthalene molecules in the crystalline state also leads to this last result.[11] For an N-fold linear array, even though the exciton band will now have N components, analogous selection rules to those illustrated in Fig. 4 still apply.[15] Obviously, the ground state is not affected by the exciton process.

Finally, in the special case of dye-molecule aggregates, one additional result of exciton-band formation is to be noted. This is the enhancement of triplet excitation accompanying molecular aggregation.[14,15] Referring to Fig. 3(b), four separate electronic processes in molecular aggregates can be delineated: (a) molecular exciton-band excitation, (b) "atomic" exciton-band excitation, (c) photoconduction excitation, and (d) triplet excitation following exciton-band excitation.[14,15] The recently reported, dramatic experiments by Albert Szent-Györgyi, demonstrating[21] great enhancement of phosphorescence in dyes frozen in water solutions, seem to be due to the foregoing phenomenon.[15]

Furthermore, since triplet states of molecules have lifetimes of a millisecond or longer, double quantum processes such as the triplet to conduction-band process shown in Fig. 3(b) may be expected, following exciton-band, then triplet excitation. The exciton-band-to-triplet process is radiationless. In most dyes, the radiationless excitation of triplets is very weak in isolated molecules.

The Förster theory,[16] with its phenomenological aspect, permits useful calculations to be made on distance of dipole-dipole transfer of excitation energy. Karreman and Steele[22] have shown that excitation energy transfer in proteins by dipole-dipole interaction through the aromatic residues is plausible according to the Förster formula.

In this section, the complications[23] of intramolecular vibrational-electronic coupling in exciton interactions have been omitted for simplicity.

5. MOLECULAR SYSTEMS AS PHOTOCONDUCTORS AND SEMICONDUCTORS

If the energy gap between the ground state (or valence band) and the conduction band of an atomic lattice is small, intrinsic electrical conductivity is observed. For example, in pure germanium and pure silicon, the energy gaps are 0.7 and 1.10 ev ($\bar{\nu}\sim$5600 and 9000 cm^{-1}), respectively.[10] In these solids, enough electrons can be thermally excited at room temperature so that dark electrical conductivity can be observed. The term semiconductor is applied,[10] arbitrarily, to solid materials which have a resistivity in the range 10^{-2} to 10^{9} ohm cm. For pure diamond, with an energy gap of 6 ev ($\bar{\nu}\sim$49 000), light of $\lambda\sim$2100 A would be required to raise an electron to a conduction band. Materials which exhibit only photoconductivity are properly classified as insulators. Since molecular ionization potentials (gas phase) are in the range[24] 8 to 10 ev, one would expect pure molecular crystals and aggregates to act as insulators, having photoconductivity perhaps for 6 to 8 ev quanta, not intrinsic electrical semiconductivity. Observations of a weak photoconductivity have been made in organic crystals absorbing near-ultraviolet light, but evidence has accumulated that, in pure single crystals, even this weak photoconductivity becomes vanishingly small as surface effects, impurities, etc., are eliminated.[25]

On the other hand, there are several reports of semiconductivity in certain organic substances in the class of molecular crystals, especially those of Eley *et al.*[26–28]

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

The general observation seems to be that, if a molecule has an optical transition at around 1 ev ($\bar{\nu}\sim$8067 cm^{-1}, $\lambda\sim$12 500 A), semiconductive behavior may be observed. This seems to be a necessary condition for semiconductivity, but not a sufficient one (see the following). For example, metal-free phthalocyanine exhibits semiconductivity (temperatures up to 400° C are used), with an indicated energy gap of ~1.2 ev ($\bar{\nu}\sim$9900 cm^{-1}, $\lambda\sim$10 200 A). However, the lowest allowed optical transition in phthalocyanine is near the infrared, but not in the infrared. The side question that arises is whether the energy gap determined by the temperature dependence of the semiconductivity is in error, or whether the lowest triplet state (which seems to be in the region of $\lambda\sim$10 000 A[29]) is involved. Contrariwise, those substances which absorb only in or near the ultraviolet, such as diphenylbutadiene and diphenyloctatetraene, exhibit no dark semiconductivity,[26–28] indicating an energy gap of greater than 2.2 ev (~50 kcal/mole, $\bar{\nu}\sim$17 000 cm^{-1}, $\lambda\sim$5800 A).

For intrinsic semiconductivity, not only must there exist a small gap between the filled levels and the unfilled levels, but there must exist sufficient intermolecular orbital overlap for a true conduction band to develop. Apparently, in the case of near ultraviolet-absorbing hydrocarbons, only a nonconducting exciton band may exist in the crystal. The stronger the van der Waals forces become, the greater is the chance that overlap interaction resulting in a conduction band will arise. Especially in the case of such dye-like molecules as pure phthalocyanines, chlorophylls, and cyanine dyes, one might anticipate significant development of a conduction band from overlap of the lowest excited-state orbitals. However, since resistivity measurements cover a wider decadic range[10] than almost any other physical property, and since conductivity measurements using powder samples are so subject to experimental artifacts, one cannot accept the observed results without caution as indicating that molecular crystals are good semiconductors at room temperature.

The semiconductivity reported[26–28] for proteins occupies a strange position. It is well known that proteins absorb in the ultraviolet, which would indicate that only photoconductivity could result, again if there were sufficient orbital overlap. Evans and Gergely,[30] in a widely misinterpreted paper, proposed that π-electron overlap *across* hydrogen bonds could give rise to sufficient overlap interaction to yield a "conduction" band. Using three different models, they obtained choices for the energy gap between the valence band (ground state) and conduction band of 3.5, 4.8, and 3.2 ev ($\lambda\sim$3500, 2600, and 3800 A), respectively. [Eley *et al.*[26–28] quote an energy gap as given by Evans and Gergely to be ~2 ev ($\lambda\sim$6200 A), but this is their value for the gap between two *filled* valence bands.] Thus, according to this discussion, and according to the pattern of results of Eley *et al.*,[26–28] *Evans and Gergely proved* (within the severe limits of their calculation) *that a pure protein should be an electrical insulator, exhibiting only photoconductivity and not semiconductivity.*

The semiconductivity observed[26–28] in proteins cannot be *intrinsic*, but if it is truly electronic, it may be *extrinsic*.[10] Extrinsic semiconductivity arises generally from the presence of lattice imperfections or impurity centers, which introduce new energy levels into the term scheme for a lattice structure. In biological systems, such perturbations of the perfect lattice probably play a very important role, and all of these discussions of the pure or perfect molecular lamellar structures are modified by their consideration. New energy levels appear in such systems, often greatly lowering the energy required for conductivity phenomena.[10]

The experiments of Arnold and Sherwood[31] demonstrate that, in the complex, heterogeneous structure of chloroplast, electron traps exist and are involved in luminescence and conductivity properties of the chloroplast. This type of investigation probably has valuable extension to other complex biological systems, and much valuable correlation with solid-state physics might be made in this connection. On the other hand, in the extremely small conductivity of weak semiconductors and insulators, it is difficult to unravel[8] ionic from electronic semiconductivity. Probably in proteins and chloroplasts, there is an appreciable ionic component of electrical conductivity.

6. PHOTOPHYSICAL STEPS IN THE PRIMARY PROCESS OF PHOTOSYNTHESIS

Since the work of Emerson and Arnold[32,33] and of Gaffron and Wohl,[34] it has been evident that absorption of energy by chlorophyll in chloroplast involves exciton behavior (called "resonance energy migration" in most of the literature). The review by Rabinowitch[20] summarized the recent status of this topic. However, since newer investigations have revealed that in pigment lamellar aggregates triplet excitation is facilitated,[14,15] the detailed primary photophysical steps require re-examination.

The lowest triplet states of chlorophyll-*a* and chlorophyll-*b* have been established spectroscopically by Becker and Kasha[29] to be at approx 1.42 ev ($\bar{\nu}\sim$11 500 cm^{-1}, $\lambda\sim$8700 A). The quantum yield of triplet excitation (via excited singlet state) is so small in monomeric chlorophyll as to suggest that the lowest triplet state could not participate in the efficient process of photosynthesis. However, the finding[14,15] that triplet excitation is greatly enhanced by aggregation of pigment molecules suggests strongly that this mechanism may play an important role in the chloroplast. The process would involve excitation to the chlorophyll exciton band, followed by triplet excitation [Fig. 3(b), first steps of process *d*]. As in other pigments,[14,15] the quantum yield of fluorescence of aggregated chlorophyll (in the leaf) is diminished many-fold over that of isolated

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

molecules in solution,[35] as would be required if enhancement of triplet excitation occurred.

The detection of photoproduced radicals in chloroplast by electron paramagnetic-resonance studies made by Commoner *et al.*[36–38] and investigated further by Calvin *et al.* (p. 157 and references 39–41) poses the question of the relation of the primary photophysical process of exciton-band excitation to the secondary photochemical process of radical production. The picture presented by Calvin and co-workers[41] proposes ionization at the exciton stage. It is possible that in the heterogeneous chloroplast structure some low-energy ionization center could exist, which traps the exciton energy, leading to the direct production of the radicals detected. But the energy of this exciton band is small (~1.8 ev, or ~41 kcal/mole), and the existence of such low-energy ionizations is unknown in molecules. Moreover, the lowest exciton state is very short-lived, somewhat shorter in life than the lowest excited singlet state of isolated chlorophyll, so that it may not be a very efficient state for a photochemical act.

The lowest triplet state of chlorophyll has been observed to have a mean lifetime of approximately 1 msec in solution,[42,43] and would be expected to have a somewhat longer lifetime in the compact chloroplast structure. It is natural to favor the triplet state for participation in the photochemical act, especially in view of the previously discussed enhancement possibility. Moreover, it is quite possible that these favorable properties of the triplet state would permit a two-quantum process to be important in the photosynthetic system. For example, as shown in Fig. 2(b), even triplet-to-conduction-band excitation might be considered. The two-quantum hypothesis of Franck[44,45] has now mainly the objection[46] that the upper state he proposed for the second quantum was a triplet state, which must have a mean lifetime of less than 10^{-12} sec.

If the lowest triplet state of chlorophyll is involved in the primary process in photosynthesis, its steady-state concentration could be too low to detect by electron paramagnetic-resonance absorption. Moreover, special orientation problems are involved which have been only recently solved.[47] It seems fairly certain that a triplet state of chlorophyll in chloroplast could not be detected by electron paramagnetic resonance.

The phenomena of solid-state physics have much to contribute to the understanding of the physical behavior of biological systems, but the richest development of experimental approaches and theoretical understanding requires that detailed attention be paid to the typically molecular behavior of the materials involved.

ADDENDUM

Since completing this manuscript, the writer has had the opportunity of studying the full text of Garrett's chapter[25] in *Semi-Conductor Chemistry*. His comprehensive review covers additional researches not included for consideration in the foregoing and emphasizes the difficulty of making significant semiconductivity measurements.

The quantitative treatment of excitons in molecular aggregates is in preparation by E. G. McRae and M. Kasha, and will be published shortly elsewhere. Reference to a preliminary communication on this work[15] has been made in the text.

ACKNOWLEDGMENTS

The author gratefully acknowledges a very useful discussion with Professor R. E. Peierls, who was Visiting Professor of Physics at the University of Colorado concurrently with the Boulder conference.

The author is indebted to Dr. C. G. B. Garrett and to Professor C. A. Hutchison, Jr., for preliminary copies of their manuscripts, referred to in the text.

BIBLIOGRAPHY

[1] Albert Szent-Györgyi, Science **93**, 609 (1941).
[2] Albert Szent-Györgyi, Nature **148**, 157 (1941).
[3] J. Frenkel, Phys. Rev. **37**, 17, 1276 (1931).
[4] J. Frenkel, Physik. Z. Sowjetunion **9**, 158 (1936) (in English).
[5] R. Peierls, Ann. Physik **13**, 905 (1932).
[6] G. H. Wannier, Phys. Rev. **52**, 191 (1937).
[7] W. R. Heller and A. Marcus, Phys. Rev. **84**, 809 (1951).
[8] Cf. F. Seitz, *The Modern Theory of Solids* (McGraw-Hill Book Company, Inc., New York, 1940).
[9] Cf. R. E. Peierls, *Quantum Theory of Solids* (Clarendon Press, Oxford, England, 1955).
[10] Cf. C. Kittel, *Introduction to Solid State Physics* (John Wiley and Sons, Inc., New York, 1956), second edition, p. 504.
[11] A. S. Davydov, Zhur. Eksptl. i Teoret. Fiz. **18**, 210 (1948). (English translation available from the present author).
[12] Cf. D. Fox and O. Schnepp, J. Chem. Phys. **23**, 767 (1955).
[13] Th. Förster, Naturwissenschaften **33**, 166 (1946).
[14] G. S. Levinson, W. T. Simpson, and W. Curtis, J. Am. Chem. Soc. **79**, 4314 (1957).
[15] E. G. McRae and M. Kasha, J. Chem. Phys. **28**, 721 (1958).
[16] Th. Förster, Ann. Physik **2**, 55 (1948).
[17] J. Franck and E. Teller, J. Chem. Phys. **6**, 861 (1938).
[18] J. Franck and R. Livingston, Revs. Modern Phys. **21**, 505 (1949).
[19] R. Livingston, J. Phys. Chem. **61**, 860 (1957).
[20] E. Rabinowitch, J. Phys. Chem. **61**, 870 (1957).
[21] Albert Szent-Györgyi, Science **124**, 873 (1956).
[22] G. Karreman and R. H. Steele, Biochim. et Biophys. Acta **25**, 280 (1957).
[23] W. T. Simpson and D. L. Peterson, J. Chem. Phys. **26**, 588 (1957).
[24] W. C. Price, Chem. Revs. **41**, 257 (1947).
[25] C. G. B. Garrett in *Semi-Conductor Chemistry* (American Chemical Society Monograph), Chap. 15 (to be published).
[26] D. D. Eley, Nature **162**, 819 (1948).
[27] D. D. Eley, G. D. Parfitt, M. J. Perry, and D. H. Taysum, Trans. Faraday Soc. **49**, 79 (1953).
[28] D. D. Eley and G. D. Parfitt, Trans. Faraday Soc. **51**, 1529 (1955).
[29] R. S. Becker and M. Kasha, J. Am. Chem. Soc. **77**, 3669 (1955).
[30] M. G. Evans and J. Gergely, Biochim. et Biophys. Acta **3**, 188 (1949).
[31] W. Arnold and H. K. Sherwood, Proc. Natl. Acad. Sci. U. S. **43**, 105 (1957).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

[32] R. Emerson and W. Arnold, J. Gen. Physiol. **15**, 391 (1932).
[33] R. Emerson and W. Arnold, J. Gen. Physiol. **16**, 191 (1932).
[34] H. Gaffron and K. Wohl, Naturwissenschaften **24**, 81, 103 (1936).
[35] P. Latimer, T. T. Bannister, and E. Rabinowitch, Science **124**, 585 (1956).
[36] B. Commoner, J. Townsend, and G. E. Pake, Nature **174**, 689 (1954).
[37] B. Commoner, J. J. Heise, and J. Townsend, Proc. Natl. Acad. Sci. U. S. **42**, 710 (1956).
[38] B. Commoner, J. J. Heise, B. B. Lippincott, R. E. Norberg, J. V. Passonneau, and J. Townsend, Science **126**, 57 (1957).
[39] M. Calvin and P. B. Sogo, Science **125**, 499 (1957).
[40] P. B. Sogo, N. G. Pon, and M. Calvin, Proc. Natl. Acad. Sci. U. S. **43**, 387 (1957).
[41] G. Tollin, P. B. Sogo, and M. Calvin, J. chim. phys. (to be published).
[42] R. Livingston and V. A. Ryan, J. Am. Chem. Soc. **75**, 2176 (1953).
[43] R. Livingston, J. Am. Chem. Soc. **77**, 2179 (1955).
[44] J. Franck, Daedalus **86**, 17 (1955).
[45] J. Franck in *Research in Photosynthesis*, H. Gaffron, editor (Interscience Publishers, Inc., New York, 1957), pp. 19–30, 142–146.
[46] G. Tollin, Arch. Biochem. Biophys. **75**, 539 (1958).
[47] C. A. Hutchison, Jr., and B. W. Mangum, J. Chem. Phys. **29**, 952 (L) (1958).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

20
General Patterns of Biochemical Synthesis

RICHARD BROOKE ROBERTS

Department of Terrestrial Magnetism, Carnegie Institution of Washington, Washington 15, D. C.

THE general pattern of biochemical synthesis is outlined in Fig. 1. The first block indicates the processes of photosynthesis which Calvin describes (p. 147). The second block represents the processes of intermediary metabolism wherein the raw materials of the medium are converted into the building blocks and energy needed for synthesis of the macromolecules. The third block represents the processes of ordering and polymerizing the intermediates. And lastly, the fourth block indicates the final folding and arranging of the macromolecules into products which may be excreted, used to maintain existing cells, or used to form new cells.

The division into these blocks is quite obvious from consideration of the growth requirements of various cells. A large variety will grow with CO_2 and light as the sole carbon and energy sources whereas others lack this capacity. Furthermore, even those cells which have this ability can hold it in abeyance and use chemical energy when the occasion demands. Accordingly, it seems quite reasonable to segregate the photosynthetic processes from the rest of the cell. Likewise, there are a number of cells which can synthesize all of their requirements from a single carbon source, such as glucose, whereas others require an exogenous supply of one or more of the more complicated molecules. In addition, most cells utilize preformed intermediates whenever they are available. Again, it seems reasonable to indicate a distinction between those processes and other operations of the cell. The division of the third and fourth stages is somewhat more arbitrary, but Kendrew's model of a protein molecule (p. 94) is quite convincing evidence that these are at least two distinct steps in the synthesis of protein.

Actual cells, in fact, do show some of the localizations of function indicated in the model cell. Calvin showed pictures of the chloroplasts which correspond to region 1, and Lehninger showed a similar localization of the Krebs-cycle reactions in the mitochondria (p. 136). Synthesis of protein and nucleic acid may occur in other special regions of the cell, such as the microsomes and the nucleus.

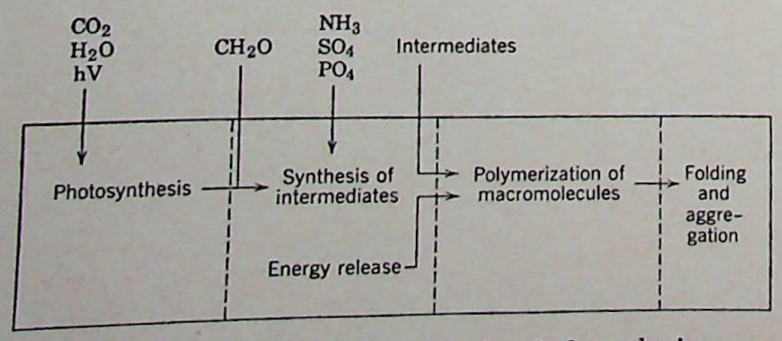


FIG. 1. General pattern of biochemical synthesis.

However, the main purpose of this diagram is not to indicate the workings of a cell, but to indicate the outline of this paper. After Calvin and Lehninger, nothing remains to be said about photosynthesis or the energy-yielding reactions. Macromolecular synthesis is covered in later papers. Thus, emphasis here is on the flow of material in the synthesis of the precursors of the macromolecules.

Before getting into the details of the synthetic pathways, there are two assigned topics which should be disposed of. The first is "least common denominators of life." One of the pleasures of biophysics is that such fascinating subjects come within the legitimate field of interest. How can life be defined? If one attempts to create life, how can success be recognized? Is the slimy precipitate in the bottom of an autoclaved mixture of organic chemicals alive or not? In part, this is a question of semantics, but the time spent in framing a definition is not wasted as it is spent in trying to select the essential from a multitude of nonessential properties.

In the model cell of the diagram, clearly photosynthesis is not essential; neither is the synthesis of intermediates, as many obviously living cells lack these capacities. Cells lacking an energy system and requiring ATP surely would be considered living. In a similar way, if some of the cells grown in tissue culture are shown to require one or more proteins, they are none the less alive! What then is essential? One obvious feature of living organisms is their capacity for growth and reproduction. Yet, these qualities alone do not define life. In framing a definition, it is necessary to choose words and concepts that include such creatures as mules which cannot reproduce and that exclude salt crystals which can grow. Furthermore, objects such as seeds and perhaps even viruses should be included. Both have the capability of catalyzing the synthesis of more of their kind even though they may not exhibit growth or other metabolic activity for long periods of time.

There presently seems to be no simple and universally acceptable answer as to how newly created life might be recognized. On the other hand, there are two particularly important properties. One is autocatalysis with its implications for growth and reproduction; another is a capacity for evolution. It would be difficult to recognize anything as alive which lacked either of these capabilities. Yet, any definition even framed in these terms is arbitrary, as the choice to include or exclude virus is largely a matter of taste.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

Another assigned topic is "comparative biochemistry at the molecular level." At this level, the processes are amazingly similar in different organisms. The reactions of glycolysis originally were worked out in yeast and in mammalian muscle; the synthesis of certain amino acids is identical in a wide variety of creatures from bacteria to man. The ribonucleoprotein particles of the microsome fraction (ribosomes for short) appear identical whether isolated from bacteria, yeast, plants, or animals. Thus, to a first approximation, all cells are the same. One can consider *E. coli* as the problem to be solved and treat all other cells as perturbations of *coli*. Thus, *Chlorella* is a variant of *coli* which has gained some additional capacities, while man is a deficient mutant which has lost a number of important synthetic pathways.

The differences between cells are often quantitative and much less frequently qualitative. The most frequently observed difference is that a pathway is completely lacking or quantitatively unimportant. It is much less common to find that two cells have found quite different solutions to a synthetic problem. Perhaps the space biologists will find some radically different mechanisms on another planet.

Consider, now, some of the details of synthetic pathways. Suppose there exists in the cell a series of reactions catalyzed by the enzymes E_1, E_2, E_3; how can these be demonstrated?

$$A \xrightarrow{E_1} B \xrightarrow{E_2} C \xrightarrow{E_3} D .$$

The most common approach is to disrupt the cells and then to isolate the enzymes and intermediates from the extract. The individual reactions can be studied then in cell-free systems. This is the classical method of biochemistry mentioned in Kornberg's introduction (p. 200). In spite of the facts that the intermediates are often present in extremely low concentrations and that the enzymes are frequently unstable, most of the important pathways of intermediary metabolism have been worked out in detail by this method. As Lehninger mentions, there are difficulties when the spatial arrangement of the enzymes in the cell is important. This method is amply illustrated in other papers (Kornberg, p. 200; Meister, p. 210; Lehninger, p. 136).

Another versatile tool is the deficient-mutant technique. Mutants which require a certain compound (*D*) for growth are isolated. Among these mutants, some are able to grow with one or another precursor (*B* or *C*) depending upon which enzyme has been lost in the mutation. The growth requirement of these different classes of mutants often suggests the reaction sequence. Such evidence also can be strengthened by demonstrating the accumulation of intermediates or the absence of a particular enzyme.

Calvin describes the kinetic approach to the identification of intermediates and his experiments demonstrate its power (p. 147). It presents certain difficulties when applied to amino-acid synthesis, because the reactions are rapid and the quantities of intermediates small. The kinetic delays, consequently, are extremely short.

FIG. 2. Radioautograph of hydrolyzed protein of cells given with C^{14} acetate as sole carbon source.

Another technique, one used extensively in our laboratory, is isotopic competition. Two different molecules, one having an isotopic label, are allowed to compete for a place in the final product. Thus, if the starting material (*A*) has a C^{14} label, C^{14} will appear in the final product (*D*). If, however, nonradioactive precursors (*B*, *C*) also are supplied, the radioactivity of the product (*D*) often is reduced strongly. Alternatively, radioactive forms of suspected precursors (*B*, *C*) may be added to compete with a nonradioactive carbon source (*A*).

The figures below show some examples of isotopic competition. Cells are grown with C^{14} acetate as the sole carbon source; they then are harvested; the protein is isolated and hydrolyzed, and the amino acids are displayed on a paper chromatogram. A radioautogram of the paper (Fig. 2) shows all of the amino acids. No competition is involved in this case and no information is gained concerning the synthetic pathways. Measurement of the different spots gives only the amino-acid content of the protein.

If C^{12} glucose also is present, the acetate enters only a limited group of amino acids (Fig. 3). Thus, a class of amino acids, which are derived more directly from acetate, is immediately segregated. If, in addition, C^{12} aspartic acid is added (Fig. 4), little radioactivity

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

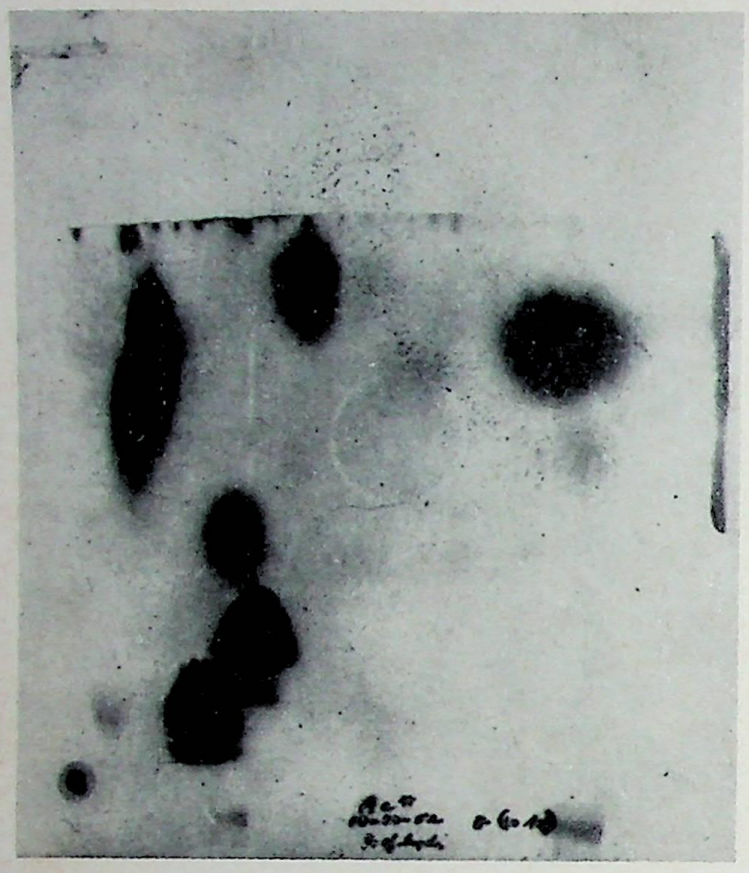


FIG. 3. Cells grown with C^{12} glucose and C^{14} acetate show radioactivity in the amino acids which incorporate acetate.

FIG. 4. C^{12} aspartic acid (in addition to C^{12} glucose and C^{14} acetate) reduces the radioactivity of aspartic acid and the amino acids derived from it.

appears in aspartic acid, threonine, methionine, isoleucine, and lysine, as these amino acids are derived from the exogenous aspartic acid. Figure 5 shows the effect of adding glutamic acid. The incorporation is

FIG. 5. C^{12} glutamic acid prevents the incorporation of acetate into the amino acids derived from the Krebs cycle leaving only leucine strongly radioactive.

suppressed in all of the amino acids except leucine. C^{12} leucine blocks the incorporation into leucine (Fig. 6) but has no effect on the other amino acids. Further details of pathways can be established by adding suspected intermediates as competitors.

This technique also gives information concerning the flow along the pathway. When radioproline is added, its sole end product is protein-bound proline. The rate of incorporation is equal to the proline content times the growth rate, as there is little if any incorporation by exchange in the growing *coli*. Glutamic acid, on the other hand, supplies carbon for proline and arginine.

A few additional experiments permit an analysis of the flow pattern in the Krebs cycle (Fig. 7). The flow of four carbon units into the cycle is equal to the flow out of the cycle into glutamic acid and its products, aspartic acid and its products, plus leakage products found in the medium. The circulation in the cycle can be calculated from the relative specific radioactivities of aspartic acid and glutamic acid when CO_2 or acetate is used as tracer. The resulting flow pattern shows that, when these cells grow on glucose, the Krebs cycle is not the main source of energy in the cell but is principally concerned with synthesis.

In contrast, when acetate is the sole carbon source, the cells must derive all of their energy from the reactions of the Krebs cycle. As a consequence, the circulation increases so that the input of acetate is roughly ten times the input of four carbon units. The reactions of the cycle must be considered quite flexible responding to changes in the environment.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

This type of analysis can be extended to include the flow of carbon into nucleic acids, lipids, and the other families of amino acids. More than 80% of the carbon incorporated can be accounted for. Minor flow patterns can be discerned, and the changes in the flow patterns caused by adaptation to other carbon sources can be observed.

Several features are notable. First, the flows are well balanced. The amino acids, although synthesized in several different systems, are each supplied at the rate needed for protein synthesis. Usually, there is very little leakage of amino acids to the medium. On occasion, alanine and valine are synthesized in excess and pour out into the culture fluid.

One clue to this balance is the control of synthesis by the product. When proline is added to the medium, even at low concentrations, the synthesis of proline promptly stops. This is not simply a reversal of the usual reactions, as the exogenous proline only is converted back to glutamic acid when present at very high concentrations. Other amino acids which strongly inhibit their own synthesis are arginine, threonine, serine, and methionine. A less complete inhibition is observed with lysine, cystine, leucine, and isoleucine.

Glutamic acid, aspartic acid, glycine, alanine, and valine, however, continue to be synthesized from glucose even though they are present in the medium. In this case, the observed competition effects must come about by a simple isotopic dilution of the internal pools of amino acids. The balance of synthesis among these

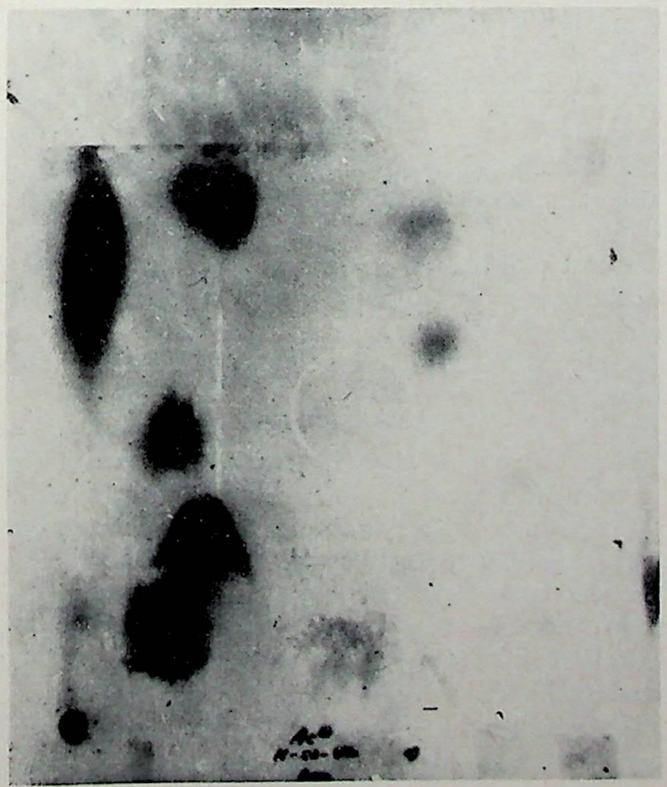


FIG. 6. C^{12} leucine supplies the carbon for leucine without influence on the radioactivity of the Krebs-cycle group of amino acids.

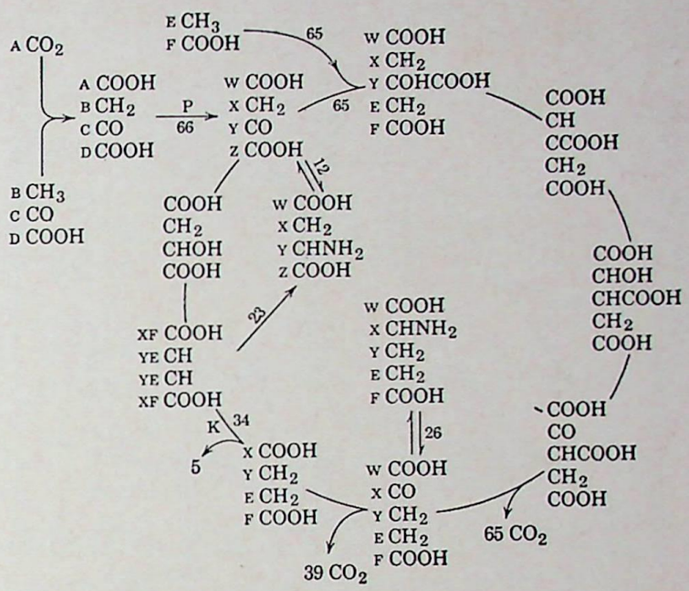


FIG. 7. Flow pattern of the Krebs cycle for cells growing on glucose as carbon and energy source. The flow numbers are expressed as micromoles per gram dry cells per 100 seconds.

and the other amino acids must be the result of other controls.

Perhaps the enzymes are present in exactly the distribution required to achieve the balance, but there may also be other less obvious control mechanisms. For example, there is a coupling between phosphorus and nitrogen metabolism which is quite obscure; external phosphate does not exchange with internal phosphate unless a nitrogen source is present. This is an unexpected result, as glycolysis continues. Growth stops when the exogenous adenine is exhausted, but the pools of adenine ribotides are not depleted.

It is clear, however, that the cell has an immediate chemical response to changes in its environment, in addition to the biological response of enzyme induction. Also, the entire system is closely coupled so that seemingly unrelated parts interact strongly.

Competition studies provide some other clues to the degree of organization of the cell. Exogenous threonine is cleaved promptly to supply the carbon of glycine. Internally synthesized threonine contributes no carbon to glycine. The enzyme which splits threonine is clearly present, yet it does not have access to the endogenous threonine. As another example, the intermediates of glycolysis do not compete with glucose. Fructose-6-phosphate, in particular, contributes no carbon to cells growing on glucose even though it supplies all of the phosphorus and completely suppresses the incorporation of orthophosphate. It will, however, compete successfully with acetate as a carbon source. These anomalies suggest highly organized enzyme systems wherein substrate molecules proceed from one active site to another, perhaps by surface diffusion.

Isotopic competition in common with the mutant technique has one basic ambiguity. It fails to distinguish whether a compound is a true intermediate or

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

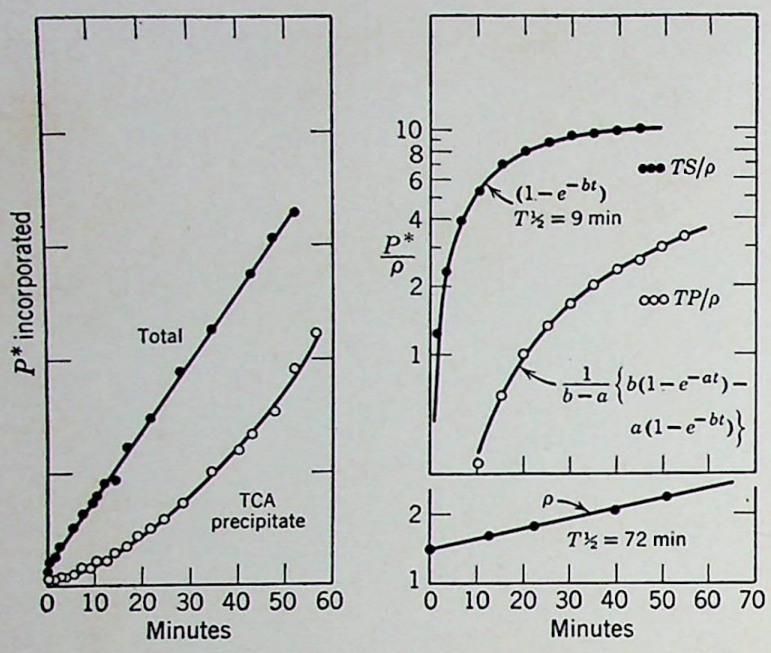


FIG. 8. The incorporation of phosphate shows transfer of P^{32} from internal pool of small TCA-soluble molecules to TCA-insoluble material. A precursor-product relationship is indicated.

merely on a side line. For example, it is not possible to tell by competition studies with C^{14} whether the keto analog of leucine is on the main line of the sequence from valine to leucine or whether it is merely a compound which is readily animated to give leucine. In this case, additional experiments using N^{15} tracers are needed.*

As a next stage in intermediary metabolism, one might expect to find the synthesis of small peptides and small polynucleotides. Unfortunately, the same methods cannot be used. Peptides are seldom found in the cells; mutants requiring peptides have not been isolated; isotopic competition fails because the peptides used as competitors are rapidly split and incorporated as individual amino acids. To investigate the reactions of amino acids and nucleotides prior to their incorporation into macromolecules, it is necessary to use kinetic methods.

In our laboratory, we have used a technique of adding a tracer compound to a growing culture and then taking a rapid succession of samples with a syringe. The cells are separated from the culture medium by filtration which requires less than two seconds. The filter can then be counted directly or extracted for chromatography.

Alternatively, the samples can be squirted directly into 5% trichloracetic acid (TCA) which extracts the small molecules. Subsequent filtration provides a sample which measures incorporation into the large molecules.

Figure 8 shows the incorporation of phosphate and its subsequent transfer from small molecules to large ones. To the accuracy of this experiment, the TCA-soluble and the TCA-precipitable material follow the curves expected for a precursor and its product. Greater accuracy and more-detailed information can be obtained by using a "pulse" of radioactivity. Figure 9 shows such an experiment; the pulse was terminated at 63 sec by adding an excess of $P^{31}O_4$. Also, the time scale was stretched out by lowering the temperature to 15° C to increase the generation time to 3.5 hours. There is a noticeable delay in the incorporation into nucleic acid after which the maximal rate is maintained for nearly 8 min. Figure 10 shows radioactivity of two nucleotides of uracil separated out by chromatography. The radioactivity of the individual phosphorus atoms of adenosine triphosphate is shown in Fig. 11. The

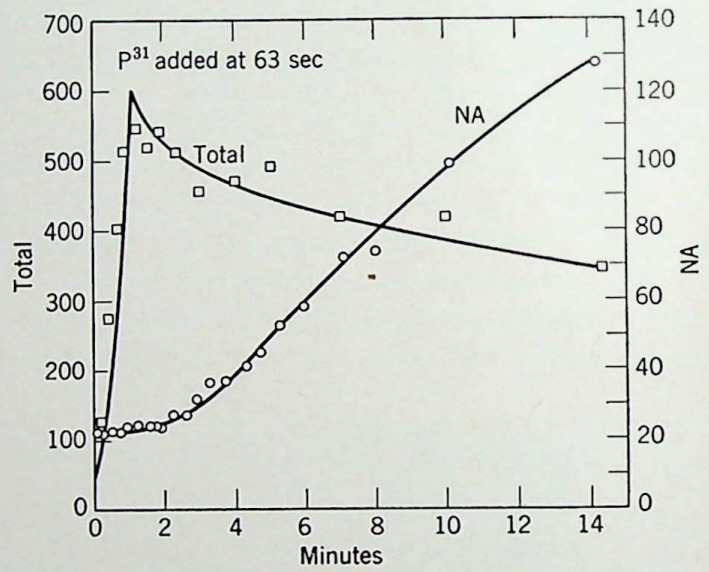


FIG. 9. Incorporation of P^{32} into total cell and nucleic-acid fraction when the radioactivity is available for a short interval.

outermost phosphorus equilibrates rapidly as it enters into the energy-transfer reactions. The innermost phosphorus appears to be the precursor of nucleic acid, as its specific radioactivity corresponds to the slope of the nucleic-acid curve.

A large number of similar experiments have been performed to follow the incorporation of amino acids into protein. As a first step, the amino acids enter the cell to form a pool wherein the concentration is typically several thousand times the concentration of the medium. This pool has extremely complicated characteristics which cannot be interpreted either in terms of simple adsorption on sites or in terms of simple transport across an impermeable membrane. A model which includes transport by a limited quantity of carrier to an array of sites, however, seems adequate for *E. coli*. In yeast, the pools are considerably more complicated.

Fig. 12 shows a typical time course of incorporation. As the size of the pool depends on the concentration in

* The experimental work described in this paper was done by the Biophysics Section, Department of Terrestrial Magnetism, Carnegie Institution of Washington, which at present includes E. T. Bolton, R. J. Britten, D. B. Cowie, and R. B. Roberts. The material up to this point is described in detail in "Studies of Biosynthesis in *E. coli*.," R. B. Roberts, P. H. Abelson, D. B. Cowie, E. T. Bolton, and R. J. Britten, Carnegie Institution of Washington Publication No. 607, 1955. A more complete account of the remainder of the experimental work can be found in the Carnegie Institution of Washington Yearbooks Nos. 54, 55, 56, and 57 (1954–1958) and in *Microsomal Particles and Protein Synthesis*, R. B. Roberts, editor (Pergamon Press, New York, 1958).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

the medium, the cells first are given a C^{12} amino acid to establish a pool of the desired size, and then the tracer is added. The pool exchanges rapidly with the external medium, as shown by the curvature of the line representing the total incorporation. Often the outflow is two-thirds the inflow. The rate of incorporation into the protein is proportional to the specific radioactivity of the pool as would be expected in a precursor-product relationship. Chromatography shows that the material extracted (either by TCA, alcohol, or water) is the unaltered amino acid and the nonextractable material is protein. Other tests showed that the amino acids incorporated during the first fifteen seconds are not concentrated at the ends of peptide chains and that they are released by acid hydrolysis at the usual rate.

These curves show that there is no appreciable quantity of partially formed proteins. If such were the case, the first protein to appear would be deficient in radio-

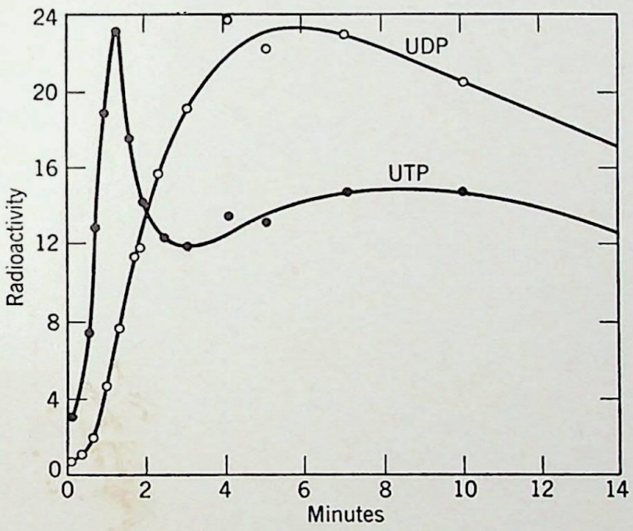


FIG. 10. Corresponding incorporation of P^{32} into uridine diphosphate and uridine triphosphate.

activity. The only kinetic delay is accounted for by the time to build up radioactivity in the pool.

Similar experiments using radiosulfate as a tracer showed that glutathione (the only peptide found in *E. coli*) is not an intermediate in protein synthesis although it can serve as a sulfur source in sulfur-deficient medium. No evidence of other peptides was found even when protein synthesis was blocked either by chloramphenical or by lack of a required amino acid.

Activated amino acids and amino acids attached to ribonucleic acid are suspected intermediates which are presently under intense study. These compounds and their possible role in protein synthesis are discussed in detail by Meister (p. 210). Unfortunately, they are present in *E. coli* in such small quantities that our methods have not given any conclusive evidence of whether or not they are intermediates of protein synthesis.

These various attempts to find intermediates beyond

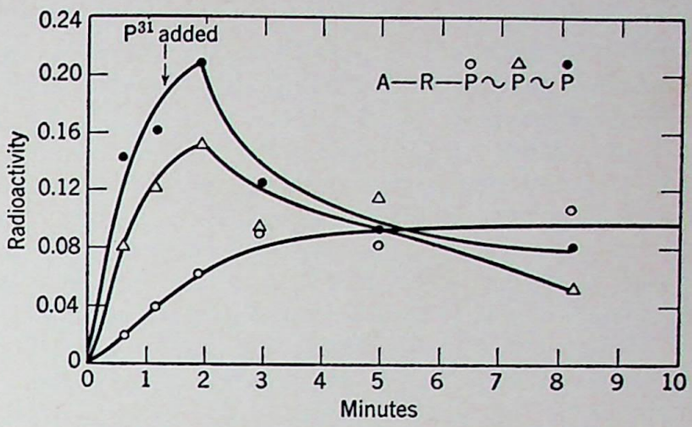


FIG. 11. Corresponding incorporation into individual phosphorus atoms of adenosine triphosphate.

the amino acids and nucleotides have failed. Looking at macromolecular synthesis from the point of view of the intermediates, one sees both amino acids and nucleotides held in the cell but exchangeable with the outside medium, until suddenly and without further preparation they appear linked into macromolecules. At this stage, it becomes necessary to determine in which macromolecules they appear. The ribosomes are of particular interest because it has been suspected for some time that RNA plays an important role in protein synthesis.

A column of Sober's modified cellulose (DEAE) provides an adequate resolution of the juice from disrupted cells. Using a salt-gradient elution, the proteins are well spread out and the protein of the ribosomes stands out as a distinct peak in the distribution. The ultraviolet absorption shows a corresponding peak with a secondary peak of nucleic acid.

Following a short exposure to radiosulfate or to radioactive amino acids, the specific radioactivity of the ribosome peak is somewhat lower than that of the other proteins. Neither is there any extra radioactivity in the proteins which spin down with the ribosomes.

Corresponding experiments with $P^{32}O_4$ show at first

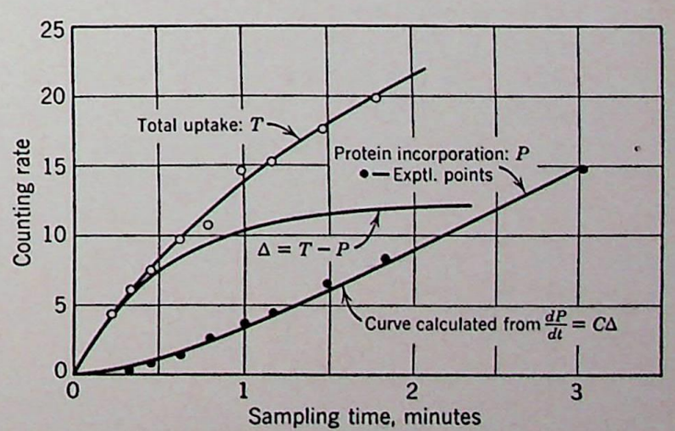


FIG. 12. Radioactive proline is rapidly incorporated by *E. coli*. The difference between the total and the TCA-insoluble (protein) measures the pool of extractable amino acid. The rate of incorporation into protein is proportional to the specific radioactivity of the pool, indicating a precursor-product relationship.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

a new and distinct peak with its radioactivity transferring only after a considerable kinetic delay to the nucleic-acid peak and later still to the ribosome peak.

These experiments, which indicate that the ribosomes are a late product of synthesis, differ from the observation in several other laboratories that radioactive amino acids appear first in the ribosomes. Two explanations for this discrepancy seem plausible. One is that in *E. coli* incorporation of amino acids by exchange is negligible whereas it may be important in mammalian cells. The prompt incorporation into the ribosomes of mammalian tissues may be owing to exchange. Another possibility is that the ribosomes of *E. coli* work much more rapidly, and there is correspondingly much less newly formed protein adhering to them. It should be pointed out also that the recovery from the columns is not complete and an important part of the kinetics thereby may be missed. In particular, the cell membranes which are lost on the column are initially high in radiophosphate.

In the space allowed, it is difficult to present more than a brief glimpse of the operation of the synthetic machinery of a cell, of the way its macromolecules are bathed in an internal medium which is different from but rapidly responding to the external environment, of the way the reactions change to meet changing conditions in the environment. It has been necessary to discuss the various elements separately, but the intercouplings and interactions are equally important to maintain a balance among the different systems. I have attempted to emphasize those findings in the study of intermediary metabolism which throw some light on the more obscure processes of protein and nucleic-acid synthesis, but they provide little information, only a few boundary conditions which must be met by any model of macromolecular synthesis.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

21
Mechanisms of Enzyme Action*

ROBERT A. ALBERTY
Department of Chemistry, University of Wisconsin, Madison 6, Wisconsin

ALMOST all reactions in living things occur spontaneously at insufficient rates and have to be catalyzed. The biological catalysts, enzymes, have all been found to be proteins. With respect to their turnover numbers, specificity, and variety of reactions catalyzed, enzymes are the most remarkable catalysts known, especially since they provide the means by which the rates of various reactions are controlled so that the rates of the large number of necessary reactions have a proper relation to each other and to the requirements of the situation.

Thus, one is presented with the tremendous challenge of trying to understand these catatalytic actions. One can distinguish roughly between two types of approaches to this problem, the kinetic approach and the chemical approach. Of course, these distinctions are arbitrary and it is really not possible to make a clear separation. Through the kinetic approach, which is emphasized in this article, one studies the effect of various independent variables on the rate of reaction and the interpretations of these effects. Through the chemical approach, one studies the effect of chemical modification of the substrate and of chemical reactions of the enzyme.

The general features of the kinetic approach have been reviewed in the literature.[1–9] For any given mechanism, it is possible to write an equation for the rate of formation of each substance involved. These equations consist of summations of concentrations or products of concentrations, each multiplied by a rate constant. These rate constants are constants under a given set of conditions, but the values of the rate constants are affected in a number of ways by components of the solution. The effects of constituents of the reaction medium may be classified according to: (a) whether they have a significant effect on the activity coefficients of reactants and the transition state; or (b) whether they are directly involved in steps in the mechanism. Effects of the first type are encountered always when the electrolyte concentration is changed. Effects of the second type are produced by hydrogen ions, metal ions, inhibitors, coenzymes, and even by buffer ions.

The solution of the simultaneous differential equations for a mechanism yields the functional form for the dependencies of the various concentrations upon time. As a further aid in this analysis, there are two types of relations between the various rate constants in a mechanism. According to the first, provided by thermodynamics, the values of the rate constants for all of the forward and reverse steps must give the correct equilibrium constant for the over-all reaction. According to the second, provided by irreversible thermodynamics, the frequency of transitions from A to B must be equal to the frequency of transitions from B to A at equilibrium, even though there are other indirect paths by which this equilibrium may be maintained. This principle of detailed balancing[10–13] applies when there are cyclic paths in a mechanism.

The kinetic behavior of a given mechanism always can be represented by a theoretical rate equation, but in the laboratory one is faced with the inverse problem of having to devise a mechanism to explain certain kinetic observations. This process requires imagination and is not one that can be summarized by a series of rules. Kinetic data show the composition of the activated complexes for the various steps in the reaction but do not prove a particular mechanism. Many conceivable mechanisms, however, may be eliminated by the use of such data.

A minimum requirement for devising a mechanism to account for kinetic data is a complete understanding of the kinetic consequences of any given mechanism. As an illustration, the simplest type of enzymatic mechanism is considered first, in which an intermediate X is formed by reaction of enzyme and substrate or of product and substrate. The concept of an intermediate complex dates back at least to 1902 when Henri and Brown independently sought to explain the fact that the velocity of inversion of sucrose by invertase is independent of the sucrose concentration at high concentrations and is directly proportional to sucrose concentration at low concentrations. In 1913, Michaelis and Menten[14] attracted considerable attention with their derivation of the initial-velocity equation assuming that the reaction $E+S \rightleftarrows X$ was in equilibrium, a rather restrictive assumption.

ENZYMATIC REACTION WITH TWO SUBSTRATES

	E	$+$	S	$\underset{k_2}{\overset{k_1}{\rightleftarrows}} X \underset{k_4}{\overset{k_3}{\rightleftarrows}}$	E	$+$	P
Initial conc.	e_0		s_0	0	e_0		0
Conc. at t	$e=e_0-x$		$s=s_0-x-p$	x	$e=e_0-x$		p

(1)

If there is more than one site per enzyme molecule, this

* This research was supported by grants from the National Science Foundation and from the Research Committee of the Graduate School of the University of Wisconsin from funds supplied by the Wisconsin Alumni Research Foundation.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

derivation will be applicable if there is no site-site interaction. If the reactants initially present are E and S, it is convenient to introduce the conservation equations ($e_0=e+x$ and $s_0=s+p+x$) and express the various concentrations at any time in terms of the instantaneous concentrations x and p of intermediate X and product P and the initial molar concentrations e_0 and s_0 of enzymatic sites and S. Owing to the conservation equations, only two of the four rate equations which may be written are independent. According to mechanism 1, the rates of change of concentration of X and P are given by

$$dx/dt=k_1se+k_4pe-(k_2+k_3)x \tag{2}$$

$$dp/dt=k_3x-k_4pe. \tag{3}$$

When these equations are written in terms of the two unknown concentrations x and p by use of the expressions under mechanism 1 and are rearranged, they become

$$dx/dt=k_1s_0e_0+(k_4-k_1)e_0p - [k_1e_0+k_1s_0+(k_4-k_1)p+k_2+k_3]x+k_1x^2 \tag{4}$$

$$dp/dt=(k_3+k_4p)x-k_4e_0p. \tag{5}$$

Transient Phase

Apparently, these simultaneous nonlinear differential equations can be solved in complete generality only with the use of a differential analyzer. The general character of the solutions is that x rises from zero and may or may not go through a maximum.[7,9,15] Families of plots of x and p for this mechanism have been given by Chance[16] for the case that $k_4=0$. Various approximate solutions have been discussed for the transient phase of the reaction when $k_4=0$.[17–20]

For the particular case that $k_1=k_4$, a condition which is met in practice at certain pH values, Eq. (4) is simplified so that it is immediately integrable to yield x as a function of time.[21] Substitution of this relation into Eq. (5) and integration yields p as a function of time. Rather than give the resulting equations, which are rather complicated, only the forms obtained when $s_0\gg e_0$ are given; they are:

$$x=\frac{k_1e_0s_0}{k_1s_0+k_2+k_3}\{1-\exp[-(k_1s_0+k_2+k_3)t]\} \tag{6}$$

$$p=\frac{k_3s_0}{k_2+k_3}\left\{1-\exp\left[-\frac{k_1(k_2+k_3)e_0t}{k_1s_0+k_2+k_3}\right]\right\} - \frac{k_1k_3s_0e_0}{(k_1s_0+k_2+k_3)}\{1-\exp[-(k_1s_0+k_2+k_3)t]\}. \tag{7}$$

Thus, for this special case, the concentration of the intermediate X rises in a first-order manner with a half-life of

$$\tau_1=0.693/(k_1s_0+k_2+k_3). \tag{8}$$

Equation (7), for the concentration of p, contains two first-order terms. The half-life for the first term is given by τ_1 and that for the second term is given by τ_2.

$$\tau_2=\frac{0.693(k_1s_0+k_2+k_3)}{k_1(k_2+k_3)e_0}. \tag{9}$$

Since $s_0\gg e_0$, $\tau_2\gg\tau_1$; that is, the half-life for the appearance of P is very large as compared with that for the appearance of the intermediate X. The second term in Eq. (7) is identical with the solution obtained below with the steady-state approximation.

If $k_1\neq k_4$ and $s_0\gg e_0$, Miller and Alberty[21] have obtained a perturbation solution which is useful if $0.1k_1<k_4<10k_1$.

The study of the transient phase of enzymatic reactions requires very sensitive and rapidly responding analytical methods. When suitable methods can be developed as Chance has done for the peroxidase and catalase reactions and for the respiratory chain,[22] more information can be obtained about the intermediates than could be obtained in general from steady-state measurements.

Steady State

If the substrate concentration is high as compared with the concentration of enzymatic sites, E must be converted to X and reconverted to E many many times before equilibrium is reached. As a result, dx/dt is very small as compared with other terms in the rate equation and, to a high degree of approximation, may be taken equal to zero. This is an approximation of considerable value for most enzyme kinetic experiments. Setting dx/dt in Eq. (4) equal to zero for the case that $s_0\gg e_0$ yields

$$x=\frac{k_4s_0+(k_1-k_4)(s_0-p)}{(k_2+k_3)+k_4s_0+(k_1-k_4)(s_0-p)}. \tag{10}$$

The concentration of the intermediate X may either increase or decrease during the steady-state phase of the reaction,[15] depending upon the sign of k_1-k_4. If $k_4>k_1$, the concentration of X increases; if $k_1>k_4$, the concentration of X decreases; and if $k_1=k_4$, the concentration of X remains sensibly constant for $t\gg\tau_1$. If the concentration of X increases in the steady state, it does not pass through a maximum; if the concentration of X decreases in the steady state, it does pass through a maximum before the steady state is established.

Substitution of Eq. (10) into Eq. (5) yields the steady-state rate equation[23] for mechanism 1.†

$$-\frac{ds}{dt}=\frac{dp}{dt}=\frac{(V_S/K_S)s-(V_P/K_P)p}{1+s/K_S+p/K_P}, \tag{11}$$

† Writing the equation in this form rather than in terms of the individual rate constants has the advantage that the constants are those directly determinable from experiment and that the steady-state rate equation for the n-intermediate case can be written in the same form.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

where

$$V_S = k_3 e_0 \qquad V_P = k_2 e_0$$
$$K_S = (k_2 + k_3)/k_1 \qquad K_P = (k_2 + k_3)/k_4. \tag{12}$$

The values of the kinetic constants usually are obtained by measuring the steady-state velocity v_f of the forward reaction immediately after the transient phase of the reaction but before the second term in the numerator and the third term in the denominator of Eq. (11) become important. Under these conditions, the velocity of the over-all reaction is given by the Michaelis-Menten equation.

$$v_f = V_S/(1 + K_S/s_0). \tag{13}$$

The initial steady-state velocity of the reverse reaction may be measured similarly if the equilibrium constant is not too large.

$$v_r = V_P/(1 + K_P/p_0). \tag{14}$$

According to this mechanism, as s_0 (or p_0) is increased, the initial steady-state velocity approaches $V_S = k_3 e_0$ (or $V_P = k_2 e_0$) which is referred to as the maximum initial velocity. If V is expressed in moles per liter per unit time and if e_0 is expressed in moles per liter, k is the turnover number of the enzyme, the number of moles of product produced per enzyme molecule (or better, per enzymatic site) per unit time. Under optimal conditions, the turnover number for catalase is greater than 10^8/min; for fumarase it is 10^5/min; and for chymotrypsin it is 0.01–10/min, depending upon the substrate.

The constants K_S and K_P in Eqs. (13) and (14) are referred to as Michaelis constants. It is evident that Michaelis constants have the units of concentration and are equal to the substrate concentration required to produce an initial steady-state velocity half as large as the maximum initial velocity. At substrate concentrations well below the Michaelis constant, the initial velocity is directly proportional to the substrate concentration.

In order to determine v_f or v_r, a very sensitive analytical method is very much needed since these velocities must generally be determined during the first several percent of reaction. Spectrophotometric, fluorometric, and titrimetric methods are those most often used, and there is a real need for better analytical methods for measuring small changes.

If the Michaelis constant for the product is small so that the third term in the denominator of Eq. (11) quickly becomes important[24] or the equilibrium constant is unfavorable, the steady-state velocity decreases rapidly during a kinetic experiment making it difficult to determine the initial velocity. To determine v_f, it is convenient to plot p/t or $-(1/t)\ln(1-p/p_{eq})$ *versus* t and extrapolate to $t=0$ to obtain v_f or v_f/p_{eq}, respectively. Such an extrapolation is linear at sufficiently short times and further kinetic information may be obtained from the initial slope.[25]

The values of the four kinetic parameters are not independent of the equilibrium constant for the over-all reaction. At equilibrium, $ds/dt = 0$, therefore, the following equation[23] is obtained from Eq. (11):

$$K_{eq} = \frac{p_{eq}}{s_{eq}} = \frac{V_S K_P}{V_P K_S} = \frac{k_1 k_3}{k_2 k_4}. \tag{15}$$

This useful test of the validity of kinetic data applies even if there are further complications in the mechanism, such as hydrogen-ion equilibria, or metal-ion binding, provided only that the effects of substrate and product concentration under a given set of conditions are represented by Eq. (11). Also, when the reaction goes essentially to completion, V_P may be calculated using this relation, provided that the other three kinetic constants have been determined and that K_{eq} is known.

Integrated Steady-State Rate Equation

Integration of Eq. (11) which is applicable if $s_0 \gg e_0$, and introduction of Eq. (15) yields[5]

$$\frac{V_S}{K_S}\left(1 + \frac{1}{K_{eq}}\right)t = \left(\frac{1}{K_S} - \frac{1}{K_P}\right)p - \left[1 + \frac{s_0}{K_P} + \frac{(1/K_S) - (1/K_P)}{1 + K_{eq}} s_0\right] \ln\left[1 - \frac{p}{p_{eq}}\right]. \tag{16}$$

If $K_{eq} \gg 1$ and $s_0 \ll K_p$, this equation becomes[26]

$$V_S t = p = K_S \ln(1 - p/s_0). \tag{17}$$

Thus, plots of $(1/t)\ln(1-p/p_{eq})$ *versus* p/t are linear. However, since it is rather complicated to calculate the maximum initial velocities and Michaelis constants from the slopes and intercepts, this approach has not been used very often.[26–28]

Once the values of K_S, V_S, K_P, and V_P have been determined, the values of the rate constants in mechanism 1 may be calculated from

$$k_1 = (V_S/e_0 + V_P/e_0)/K_S, \tag{18}$$
$$k_2 = V_P/e_0, \tag{19}$$
$$k_3 = V_S/e_0, \tag{20}$$
$$k_4 = (V_S/e_0 + V_P/e_0)/K_P. \tag{21}$$

Thus, from steady-state kinetic studies, it is possible to obtain the values of all four of the rate constants in mechanism 1. If the number n of enzymatic sites per molecule is unknown, the calculated values of k_1 to k_4, using the molar concentration of the enzyme in Eqs. (18) to (21), will be larger than the true constants by an integer factor n.

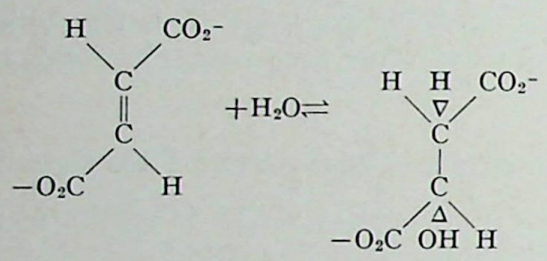
Although mechanism 1 has certain features which are

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

common to many enzymatic reactions, it can be stated confidently that Eq. (2) *is too simple to represent the mechanism of any enzymatic reaction*. However, this mechanism is of great importance for two reasons. First, it can be extended in many ways to include further steps. Second, under a given set of conditions, reactions following more-complicated mechanisms may appear to follow mechanism 1. Thus, it is often useful to calculate the magnitudes of the four apparent rate constants. As an example of the extension of the simple mechanism, the mechanism of the fumarase reaction is considered.

Effect of *p*H Changes on the Fumarase Reaction

Fumarase catalyzes the hydration of fumarate (F) to L-malate (M). The effect of *p*H changes on the



(22)

maximum initial velocities and Michaelis constants for the forward and reverse reactions can be accounted for in terms of the mechanism[29] shown in Fig. 1. Two intermediates are required because the maximum initial velocities for the forward and reverse reactions have different *p*H optima.

As expected for the steady-state treatment of a mechanism of this type, only the equilibrium constants for the acid dissociations (represented by K) are involved and not the individual rate constants which make up these equilibrium constants. The steady-state rate equations may be arranged in the form of Eqs. (13) and (14).[15] The Michaelis constants and maximum velocities are given by

$$V_F=\frac{V_F'e_0}{1+(\mathrm{H}^+)/K_{aEF}'+K_{bEF}'/(\mathrm{H}^+)} \tag{23}$$

$$K_F=K_F'\frac{1+(\mathrm{H}^+)/K_{aE}+K_{bE}/(\mathrm{H}^+)}{1+(\mathrm{H}^+)/K_{aEF}'+K_{bEF}'/(\mathrm{H}^+)} \tag{24}$$

$$V_M=\frac{V_M'e_0}{1+(\mathrm{H}^+)/K_{aEM}'+K_{bEM}'/(\mathrm{H}^+)} \tag{25}$$

$$K_M=K_M'\frac{1+(\mathrm{H}^+)/K_{aE}+K_{bE}/(\mathrm{H}^+)}{1+(\mathrm{H}^+)/K_{aEM}'+K_{bEM}'/(\mathrm{H}^+)}. \tag{26}$$

The ten kinetic parameters in these four equations represent the experimental data over the range *p*H 5 to 8.5, and they have been determined at five ionic strength values at 25°C using tris-(hydroxymethyl)-aminomethane acetate buffers. Only two of them, K_{aE} and K_{bE}, can be identified with constants in the mechanism. The other eight kinetic parameters are related to the constants in the mechanisms by

$$V_F'=\frac{k_3k_5}{k_3+k_4+k_5} \tag{27}$$

$$K_F'=\frac{k_2k_5+k_2k_4+k_3k_5}{k_1(k_3+k_4+k_5)} \tag{28}$$

$$V_M'=\frac{k_2k_4}{k_2+k_3+k_4} \tag{29}$$

$$K_M'=\frac{k_2k_5+k_2k_4+k_3k_5}{k_6(k_2+k_3+k_4)} \tag{30}$$

$$K_{aEF}'=\frac{k_3+k_4+k_5}{(k_4+k_5)/K_{aEF}+k_3/K_{aEM}} \tag{31}$$

$$K_{bEF}'=\frac{(k_4+k_5)K_{bEF}+k_3K_{bEM}}{k_3+k_4+k_5} \tag{32}$$

$$K_{aEM}'=\frac{k_2+k_3+k_4}{(k_2+k_3)/K_{aEM}+k_4/K_{aEF}} \tag{33}$$

$$K_{bEM}'=\frac{(k_2+k_3)K_{bEM}+k_4K_{bEF}}{k_2+k_3+k_4}. \tag{34}$$

The K_a' and K_b' values are calculated from the plots of V *versus* *p*H and V/K *versus* *p*H.[30] The eight relations of Eqs. (27) to (34), however, are insufficient to determine the values of the other ten constants for the mechanism illustrated in Fig. 1. But, these relations do place considerable restrictions on the values of the various rate constants and acid-dissociation constants. A detailed analysis[15] of the data shows that k_3 and k_4 may be determined within their experimental uncertainties, and *minimum* values are obtained for k_1, k_2, k_5, and k_6.

The experimental data show that the minimum values of the rate constants for the bimolecular reaction of the substrate with the enzymatic site are of the

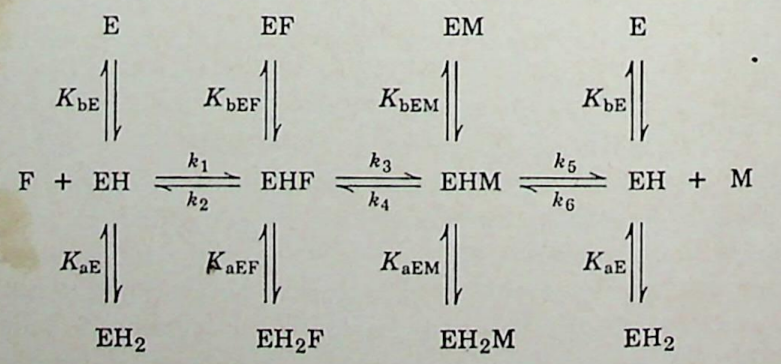


FIG. 1. Mechanism of the fumarase reaction required to explain the effect of *p*H on the kinetics of the forward and reverse reactions.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

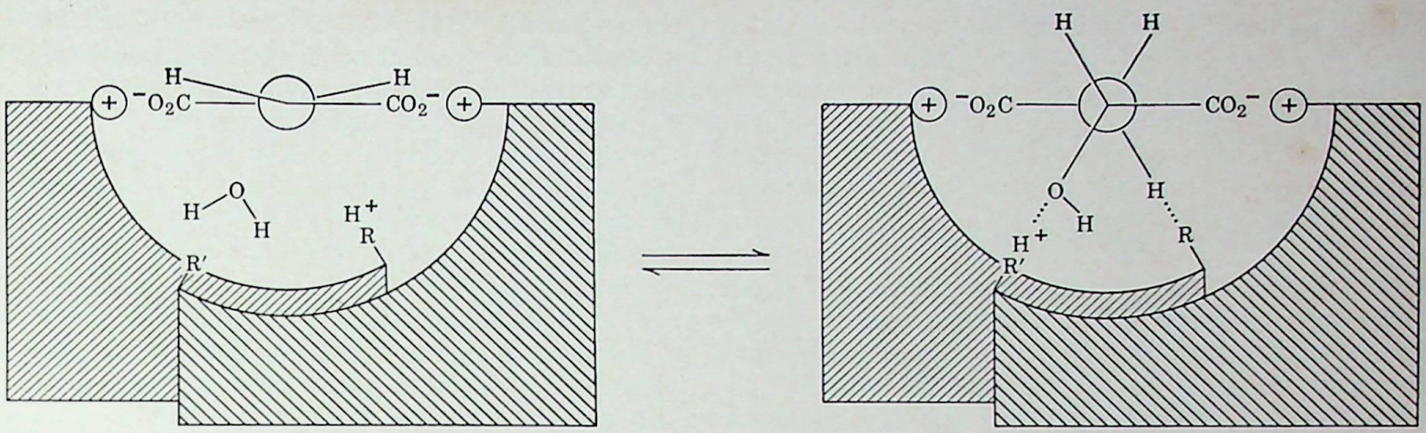


FIG. 2. Schematic representations of the enzyme-fumarate and enzyme-L-malate complexes indicating the two weak acidic groups R and R′ in the catalytic site and accounting for the *cis*-hydration [from T. C. Farrar, H. S. Gutowsky, R. A. Alberty, and W. G. Miller, J. Am. Chem. Soc. **79**, 3978 (1957)].

order of 10^9 liter mole^{-1} sec^{-1} when extrapolated to zero ionic strength. This high value for the rate constant of a second-order reaction in solution indicates that the reaction rate is controlled by diffusion. The maximum rate with which the substrate could diffuse into a hemispherical sink on a flat surface of a protein molecule is expressed as a second-order rate constant (in liter mole^{-1} sec^{-1}) by

$$k=\frac{2\pi N}{1000}R_{12}D_{12}f, \tag{35}$$

where N is Avogadro's number, R_{12} is the reaction radius (sum of the radii of site and substrate), D_{12} is the sum of the diffusion coefficients of substrate and enzyme, and f is an electrostatic factor which corrects for the slowing down or speeding up of the diffusion process owing to electrostatic interactions between the substrate and enzymatic site.[31] The theoretical values of the second-order rate constant calculated for fumarase using Eq. (35) are greater than the minimum value obtained from the kinetics by a factor of 3 at most.[32] Therefore, it is likely that the first step in the enzymatic reaction is simply the formation of an ion pair which subsequently undergoes an intermolecular reaction. The effect on the second-order rate constant of changing the ionic strength indicates that the enzymatic site has two positive charges.[32]

Because of the high value of this second-order rate constant, the half-time for the rise in concentration of intermediates in the fumarase reaction is of the order of 10^{-5} to 10^{-4} sec. Such rapid changes cannot be studied by mixing methods, but may be measured by relaxation methods such as those developed by Eigen.[33]

Our interpretation of the total effect of two acid groups in the enzymatic site is that these groups donate and accept protons in the catalytic process. This is illustrated in Fig. 2 which shows the ionizable groups on one side of the bound substrate ion as required by the stereochemistry of the deuteromalic acid obtained when the reaction is carried out in D_2O.[34] According to this mechanism, the enzyme is a clever acid-base catalyst which makes available a proton at exactly the required position in space and which has a site to accept a proton at precisely the right position. This sort of mechanism accounts for the absolute stereochemical specificity.

Inhibition and Activation

Many substances combine with enzymes in such a way as to reduce or even to increase the steady-state velocity. Compounds which are structurally similar to the substrate often combine at the enzymatic site. This may be represented by adding the reaction

$$E+I=EI, \tag{36}$$

where EI cannot combine with substrate and I is the inhibitor. The steady-state rate equation for reactions (1) and (36) becomes

$$v_f=\frac{V_s}{1+K_S(1+i/K_I)/s}, \tag{37}$$

where i is the concentration of the inhibitor I and K_I is the dissociation constant for the EI complex. An inhibitor whose effect on the rate may be represented in this way is called a competitive inhibitor. Information on the competitive-inhibition constants for a number of inhibitors is useful in learning about the nature of the enzymatic site. However, since inhibition constants will, in general, depend upon the *p*H, what is needed in the study of inhibitors are kinetic data over a range of *p*H, so that the affinity of a *particular* ionized form of the enzymatic site for the inhibitor may be calculated.[35]

If the inhibitor combines only at a site where substrate does not combine, the effect on the kinetics is not competitive and the inhibitor concentration enters the rate equation in a different way. A mixture of competitive and noncompetitive effects may be accounted

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

for by a mechanism of the type[5]

$$\begin{gathered} E+S \underset{k_2}{\overset{k_1}{\rightleftharpoons}} ES \overset{k_1}{\rightarrow} E+P \\ E+I \underset{k_5}{\overset{k_4}{\rightleftharpoons}} EI \\ E+I \underset{k_7}{\overset{k_6}{\rightleftharpoons}} IE \\ IE+S \underset{k_9}{\overset{k_8}{\rightleftharpoons}} IES \overset{k_{10}}{\rightarrow} IE+P \\ EI+I \underset{k_{12}}{\overset{k_{11}}{\rightleftharpoons}} IEI, \end{gathered} \tag{38}$$

where writing I to the left of E means that it is combined at a neighboring site. The combination of an ion at a neighboring site may increase the rate of reaction of the enzyme-substrate intermediate to yield product. The steady-state rate equation for this mechanism is the Michaelis equation with

$$V=k_3(E)_0\frac{1+k_{10}K_{ES}i/k_3K_{IES}K_{IE}}{1+K_{ES}i/K_{IES}K_{IE}} \tag{39}$$

$$K=K_{ES}\frac{1+i(1/K_{EI}+1/K_{IE})+i^2/K_{EI}K_{IEI}}{1+K_{ES}i/K_{IES}K_{IE}}, \tag{40}$$

where $K_{ES}=(k_2+k_3)/k_1$, $K_{IES}=(k_9+k_{10})/k_8$, $K_{EI}=k_5/k_4$, $K_{IE}=k_7/k_6$, and $K_{IEI}=k_{12}/k_{11}$. Accordingly, linear plots of $1/v$ *versus* $1/s$ are obtained, but both the slope and intercept vary with the concentration of inhibitor in this general case. Mechanisms with further steps have been discussed in detail in the literature.[36]

Buffer ions, and even substrate ions, may cause activation or inhibition as discussed above for substance I.[37] If substrate is bound at a neighboring site and this alters the rate constants in the catalytic mechanism, the steady-state kinetics will not follow the Michaelis-Menten rate equation except in the limit of low substrate concentrations. The binding of substrate at higher substrate concentrations may cause either inhibition or activation, effects which are not provided for by the simple Michaelis-Menten mechanism.

ENZYMATIC REACTIONS WITH FOUR SUBSTRATES

Such a reaction may be represented in general by

$$A+B=C+D. \tag{41}$$

An example is the alcohol-dehydrogenase reaction

$$\text{alcohol}+\text{DPN}=\text{aldehyde}+\text{DPNH}+\text{H}^+, \tag{42}$$

where DPN is diphosphopyridine nucleotide (a coenzyme) and DPNH is the reduced form. This reaction is catalyzed by the enzyme alcohol dehydrogenase. The kinetics of this reaction have been extensively investigated by Theorell and co-workers.[38–41]

Steady-state treatments have been given for a number of simple mechanisms that can be conceived of for a reaction of this type.[42] The objective of the kinetic experiments is to determine which of the paths is followed and which are the important intermediates and rate-determining steps. The same type of rate law is obtained for several of these mechanisms, and, as shown by Dalziel,[43] it is convienient to write the steady-state rate equations for the initial velocity when A and B are mixed with enzyme in the general form

$$\frac{e_0}{v}=\phi_0+\frac{\phi_1}{a}+\frac{\phi_2}{b}+\frac{\phi_{12}}{ab} \tag{43}$$

wherever this is possible. Here a and b represent the initial concentrations of A and B, and ϕ_0, ϕ_1, ϕ_2, and ϕ_{12} are parameters which may be obtained readily from the experimental data by making plots of $1/v$ *versus* $1/a$ with b held constant and $1/v$ *versus* $1/b$ with a held constant. There are certain relationships between these parameters for certain mechanisms and various relations between the parameters for the forward and reverse reactions. This is illustrated for the simplest type of mechanism only. Both steady-state and transient-state kinetic data consistent with the following mechanism have been obtained[38–41] for the alcohol-dehydrogenase reaction

$$\begin{gathered} E+A \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} EA \\ EA+B \underset{k_{-2}}{\overset{k_2}{\rightleftharpoons}} EC+D \\ EC \underset{k_{-3}}{\overset{k_3}{\rightleftharpoons}} E+C. \end{gathered} \tag{44}$$

The steady-state rate equations for the forward and reverse reaction are conveniently written in the forms

$$\frac{e_0}{v_f}=\left(\frac{1}{k_3}+\frac{1}{k_1a}\right)+\frac{1}{k_2b}\left(1+\frac{k_{-1}}{k_1a}\right), \tag{45}$$

$$\frac{e_0}{v_r}=\frac{1}{k_{-1}}+\frac{1}{k_{-3}c}+\frac{1}{k_{-2}d}+\frac{k_3}{k_{-2}k_{-3}cd}. \tag{46}$$

It can be seen that, from studies of the forward reaction, the values of k_1, k_2, k_3, and k_{-1} may be obtained, and that, from the reverse reaction, the values of k_{-1}, k_{-2}, k_{-3}, and k_3. Thus, all six of the rate constants are obtained and check values are obtained for k_3 and k_{-1}. There is a further relation between these kinetic

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

parameters which must be obeyed; namely,

$$K_{eq}=\frac{c_{eq}d_{eq}}{a_{eq}b_{eq}}=\frac{k_1k_2k_3}{k_{-1}k_{-2}k_{-3}}. \quad (47)$$

Further checks may be obtained by the use of experiments in which one of the products is added along with the two reactants.[44] For example, if A, B, and C are added, the steady-state rate equation becomes

$$\frac{(E)_0}{v_f}=\frac{1}{k_3}+\frac{1}{k_1a}\left(1+\frac{k_{-3}c}{k_3}\right)+\frac{1}{k_2b}+\frac{k_{-1}}{k_1k_2ab}\left(1+\frac{k_{-3}c}{k_3}\right). \quad (48)$$

Thus, by determining the effect of C on the initial steady-state velocity, it is possible to obtain k_{-3}/k_3 and, consequently, a value of k_{-3} from the forward reaction, since k_3 is known. By adding D, it is possible to obtain k_{-2}. Thus, all six of the rate constants can be obtained from steady-state velocities *of the forward reaction only*.

If there is a ternary complex in the second step of mechanism 44, there are eight rate constants in the mechanism, and all eight may be determined by means such as those just described. Actually, there must be a ternary complex in this mechanism, since a proton is transferred stereochemically specifically from DPNH to aldehyde to form an optically active monodeuteroethanol, but apparently its steady-state concentration is very low under the usual experimental conditions.

Hearon[7] has given a generalized mathematical treatment of mechanisms of reactions of this type which provides for the possibility of ternary complexes and for further reactants and products. Some rather unusual kinetic features are encountered in the steady-state studies of coenzyme mechanisms.[24]

USES OF ISOTOPES IN ELUCIDATING MECHANISMS

The use of isotopes often permits the determination of the cleavage point in an hydrolysis reaction, as represented by[45,46]

$$AOB+H_2O^*=AO^*H+BOH \quad \text{or} \quad AOH+BO^*H, \quad (49)$$

or of the position of the hydrogen atom transferred, as in

$$CH_3CD_2OH+DPN^+=CH_3CDO+\overset{*}{D}PND. \quad (50)$$

Isotope-exchange experiments can be used to distinguish between the following two mechanisms for a transfer reaction.

$$AB+D=A+BD \quad (51)$$

(I)
$$AB+E=EAB$$
$$EAB+D=EABD=EBD+A$$
$$EBD=E+BD$$

(II)
$$AB+E=BE+A$$
$$D+BE=BD+E$$

If the mechanism is given by (II), a small amount of free A is formed when AB is added to the enzyme, even when D is absent. If labeled A is placed in the medium, the label appears in AB. If the mechanism is given by (I), no incorporation of label into AB occurs unless D is present also.

When the fumarase reaction is carried out in D_2O, a specific monodeutero-L-malate is formed. The absence of an isotope effect in the enzymatic dehydration of this monodeutero-L-malate suggested that the deuterium might be incorporated in a rapid reversible step. In order to determine whether such an exchange precedes the rate-determining step in the fumarase reaction, an experiment was carried out with L-malate as substrate in D_2O, and L-malate was isolated at various times and analyzed for deuterium.[47] The experiment was carried out under conditions where the Michaelis constants for malate and fumarate were equal so that the over-all reaction was pseudo first order, and the incorporation expected in the absence of exchange could be calculated. The experimental data were in agreement with this calculation showing that there is no direct exchange of the specific proton of malate with the medium.

Investigations of the type discussed in the foregoing reveal the nature of the kinetically important steps in the reaction and the values of the rate constants, but they do not reveal much about the chemical nature of these intermediates. A basic difficulty in understanding the chemistry of the intermediate compounds at the present time is the lack of information about the chemical structure of the enzymatic site. Progress is being made, however, and recent gains in the understanding of the chemical action of chymotrypsin (Neurath, p. 185) is an example, where the chemical mechanism has been filled into these blank spaces in the formal mechanism.

BIBLIOGRAPHY

[1] B. Chance, Advances in Enzymol. **12**, 153 (1951).

[2] J. Z. Hearon, Physiol. Rev. **32**, 499 (1952).

[3] W. D. McElroy and B. Glass, editors, *The Mechanism of Enzyme Action* (The Johns Hopkins Press, Baltimore, Maryland, 1954).

[4] *The Physical Chemistry of Enzymes*, Discussions Faraday Soc. **20** (1956).

[5] R. A. Alberty, Advances in Enzymol. **17**, 1 (1956).

[6] M. Dixon and E. C. Webb, *Enzymes* (Academic Press, Inc., New York, 1958), p. 62.

[7] J. Z. Hearon, S. A. Bernhard, S. Freiss, D. J. Botts, and M. F. Morales in *The Enzymes*, P. D. Boyer, H. Lardy, and K. Myrbäck, editors (Academic Press, Inc., New York, 1958), second edition (to be published).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

[8] R. Lumry in *The Enzymes*, P. D. Boyer, H. Lardy, and K. Myrbäck, editors (Academic Press, Inc., New York, 1958), second edition (to be published).
[9] R. A. Alberty in *The Enzymes*, P. D. Boyer, H. Lardy, and K. Myrbäck, editors (Academic Press, Inc., New York, 1958), second edition (to be published).
[10] R. H. Fowler, *Statistical Mechanics* (Cambridge University Press, Cambridge, England, 1929).
[11] G. N. Lewis, Proc. Natl. Acad. Sci. U. S. **11**, 179 (1925).
[12] L. Onsager, Phys. Rev. **37**, 405 (1931).
[13] J. S. Thomsen, Phys. Rev. **91**, 1263 (1953).
[14] L. Michaelis and M. L. Menten, Biochem. Z. **49**, 333 (1913).
[15] R. A. Alberty and W. H. Peirce, J. Am. Chem. Soc. **79**, 1526 (1957).
[16] B. Chance, J. Biol. Chem. **151**, 553 (1943).
[17] M. F. Morales and D. E. Goldman, J. Am. Chem. Soc. **77**, 6069 (1955).
[18] H. Gutfreund, Discussions Faraday Soc. **20**, 167 (1955).
[19] K. J. Laidler, Can. J. Chem. **33**, 1614 (1955).
[20] P. A. T. Swoboda, Biochim. et Biophys. Acta **23**, 70 (1957).
[21] W. G. Miller and R. A. Alberty, J. Am. Chem. Soc. **80**, 5146 (1958).
[22] B. Chance and G. R. Williams, Advances in Enzymol. **17**, 65 (1956).
[23] J. B. S. Haldane, *Enzymes* (Longmans, Green and Company, London, 1930).
[24] E. L. King, J. Phys. Chem. **60**, 1378 (1956).
[25] R. A. Alberty and B. M. Koerber, J. Am. Chem. Soc. **79**, 6379 (1957).
[26] A. C. Walker and C. L. A. Schmidt, Arch. Biochem. **5**, 445 (1944).
[27] T. H. Applewhite and C. Niemann, J. Am. Chem. Soc. **77**, 4923 (1955).
[28] K. A. Booman and C. Niemann, J. Am. Chem. Soc. **77**, 5733 (1955).
[29] C. Frieden and R. A. Alberty, J. Biol. Chem. **212**, 859 (1955).
[30] R. A. Alberty and V. Massey, Biochim. et Biophys. Acta **13**, 347 (1954).
[31] P. Debye, Trans. Am. Electrochem. Soc. **82**, 265 (1942).
[32] R. A. Alberty and G. G. Hammes, J. Phys. Chem. **62**, 154 (1958), and forthcoming publication.
[33] M. Eigen, Discussions Faraday Soc. **17**, 194 (1954).
[34] T. C. Farrar, H. S. Gutowsky, R. A. Alberty, and W. G. Miller, J. Am. Chem. Soc. **79**, 3978 (1957).
[35] R. A. Alberty in *Molecular Structure and Biological Specificity*, L. Pauling and H. A. Itano, editors (American Institute of Biological Sciences, New York, 1957).
[36] J. Botts and M. F. Morales, Trans. Faraday Soc. **49**, 696 (1953).
[37] R. A. Alberty and R. M. Bock, Proc. Natl. Acad. Sci. U. S. **39**, 895 (1953).
[38] H. Theorell and R. Bonnichsen, Acta Chem. Scand. **5**, 1105 (1951).
[39] H. Theorell and B. Chance, Acta Chem. Scand. **5**, 1127 (1951).
[40] H. Theorell, A. P. Nygaard, and R. Bonnichsen, Acta Chem. Scand. **8**, 1490 (1954).
[41] H. Theorell, A. P. Nygaard, and R. Bonnichsen, Acta Chem. Scand. **9**, 1148 (1955).
[42] R. A. Alberty, J. Am. Chem. Soc. **75**, 1928 (1953).
[43] K. Dalziel, Acta Chem. Scand. **11**, 1706 (1957).
[44] R. A. Alberty, J. Am. Chem. Soc. **80**, 1777 (1958).
[45] D. E. Koshland and S. S. Stein, J. Biol. Chem. **208**, 139 (1954).
[46] H. F. Fisher, E. E. Conn, B. Vennesland, and F. H. Westheimer, J. Biol. Chem. **202**, 687 (1953).
[47] R. A. Alberty, W. G. Miller, and H. F. Fisher, J. Am. Chem. Soc. **79**, 3973 (1957).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

22
Protein Structure and Enzyme Action

HANS NEURATH

Department of Biochemistry, University of Washington, Seattle 5, Washington

THE idea that proteins are enzymes is so firmly entrenched in our minds that it may be regarded as a "self-evident truth." However, only a few decades ago, this concept was the subject of a lively controversy, and the first report, by J. B. Sumner,[1] of the crystallization of a protein having all of the properties of the enzyme urease, was greeted with critical reservations. The subsequent crystallization of other enzymes, including several proteolytic enzymes, notably by Northrop, Kunitz, and Herriott,[2] has added support to the identification of enzymes as proteins, and today over 130 enzymes of known specificity and function have been obtained in crystalline form (see Dixon and Webb[3]).

Of course, crystallinity, *per se*, is neither a necessary nor a sufficient criterion for the protein nature of a compound. Suffice it to say that every criterion of protein chemistry has been found applicable to purified enzymes, and no enzyme has yet been found which would be an exception to this statement. Thus, it has been known for many years that processes leading to protein denaturation, in turn, are likely to cause enzyme inactivation, and, when denaturation could be reversed, enzyme activity would reappear. Typical examples for this statement are the work of Kunitz,[4] on the heat denaturation of the soy-bean trypsin inhibitor, and the work of Eisenberg and Schwert,[5] on the reversible denaturation of chymotrypsinogen. As denaturation may be described as an effect on the secondary and tertiary structure of proteins, it may be said, therefore, that enzymatic activity is related to a specific configuration of the protein molecule or of a part thereof. The enzymatic activity is evidently also related to the primary structure of the protein molecule; i.e., the sequence of amino acids along the polypeptide chain but not all amino-acid side chains are required for activity. Thus, in some cases, some of the side chains may be chemically modified without adversely affecting catalytic functions, and in others, some portions of the polypeptide chain may be split off without deleterious effects on activity. Furthermore, enzymes of like specificity but of different origin have been found to differ from one another in amino-acid composition. It is evident, therefore, from these sketchy comments, that not the entire molecular structure of the protein is necessarily involved in the enzymatic process, a concept which is fully consistent, on the molecular level, with the idea previously held by enzymologists that an enzyme has an "active site" or "active center" which is the primary locus of the catalytic process. Such a view, though probably an oversimplification, provides a convenient starting point for an inquiry into the relation of protein structure to enzymatic function.

It seems desirable at this point to differentiate between the catalytic activity as such and the specificity of the enzyme, because, unlike inorganic catalysts or catalysts of simpler molecular structure, enzymes are endowed with a high degree of substrate specificity. Though the same type of bond may be hydrolyzed by an enzyme in different substrates, the chemical environment of the bond being split will determine the specificity requirements of an enzyme. Consider, as an example, the proteolytic enzymes which, as the name implies, catalyze the hydrolysis of proteins. A protein may be considered a repeating pattern of peptide bonds, but each peptide bond will find itself in a different environment, depending on the nature of amino acids contributing the CO and NH groups. Table I indicates the specificity requirements of trypsin, chymotrypsin, and pepsin.

This specificity is related to the binding of substrate by the enzyme, the simplest postulate being that the enzyme has a structure complementary to that of the substrate. Conversely, if the enzyme is presented with a typical substrate, the bond being hydrolyzed can be varied within relatively wide limits provided that the various bonds do not differ greatly in bond strength.

TABLE I. Peptide bonds hydrolyzed by proteolytic enzymes.

Enzyme	Substrate	Bonds hydrolyzed
Trypsin	Insulin (oxidized *B*) Ribonuclease (oxidized)	—Arg—Gly—; —Lys—Ala— —Lys—Phe—; —Lys—Ser—; —Arg—Asp(NH_2)—; —Lys—Asp—; —Lys—Asp(NH_2)—; —Lys—His—.
Chymotrypsin	Insulin (oxidized *B*)	—Tyr—Leu—; —Phe—Tyr—; —Tyr—Thr—; (—Leu—Tyr—)
	Melanophore-stimulating hormone	—Tyr—Lys—; —Phe—Arg—; Try—Gly—.
Pepsin	β-Corticotropin	—Glu—Asp—; —Glu(NH_2)—Leu—; —Leu—Ala—; —Phe—Pro—.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

TABLE II. Types of bonds hydrolyzed by chymotrypsin.

Peptide	$R_1CONH-CHR_2-CO-\nearrow-NH-$
Ester	$R_1CONH-CHR_2-CO-\nearrow-OR_3$
Amide	$R_1CONH-CHR_2-CO-\nearrow-NH_2$
Hydroxamide	$R_1CONH-CHR_2-CO-\nearrow-NHOH$
Hydrazide	$R_1CONH-CHR_2-CO-\nearrow-NHNH_2$
−C−C−	$C_6H_5OH-CH_2-CH_2-CO-\nearrow-CH_2-COOC_2H_5$

Thus, chymotrypsin can hydrolyze not only peptide bonds but also amide, ester, hydroxamide, hydrazide, and certain labile C−C bonds (Table II). The feature common to all these substrates is that the surroundings of the bonds are similar and meet the specificity requirements of the enzyme.

Having thus proposed the existence of a catalytic site and a specificity site within the enzyme molecule, one may start the inquiry by posing the following questions:

(1) What is the chemical configuration of the active site responsible for catalysis?

(2) Is the same or a different site responsible for the specificity of the enzyme, and what is the chemical nature of that site?

(3) How many active sites (including both the catalytic and the specificity sites) exist in an enzyme molecule?

(4) What is the function of the remainder of the mass of the enzyme molecule, or, to put this question in an operational frame, how extensively can an enzyme be degraded before it loses its enzymatic activity?

At this point, one should differentiate clearly between the protein portion of the enzyme and the nonprotein portion, which may also be part of the active site. The nonprotein moiety, usually referred to as a prosthetic group, may be an organic molecule of a different kind such as the flavin group in a flavoprotein, it may be a metal ion, or it may be a combination of an organic molecule and a metal ion. Such a prosthetic group and the particular group of the protein involved in the specific interaction are operationally included in the definition of the active site. The important role of metals in these catalytic processes is discussed in recent reviews by Vallee[6] and by Smith *et al.*[7] which should be consulted.

Consider now enzymes which, as far as is known, do not contain a prosthetic group, and where the active site seems to be composed entirely and exclusively of specific amino-acid residues. Such enzymes include certain, but by no means all, proteolytic enzymes. Trypsin, chymotrypsin, and papain are proteolytic enzymes which are believed to be free of nonprotein prosthetic groups, whereas carboxypeptidase is definitely known to be a zinc-metalloenzyme[8] and is not to be considered here. The three proteolytic enzymes mentioned in the foregoing have each been fully characterized with respect to molecular weight and certain other physicochemical properties,[9,10] some of them with respect to their amino-acid composition and partial amino-acid sequences.*,[11] The physicochemical properties of an inactive precursor† of one of these enzymes, chymotrypsinogen, are summarized in Table III. It is clear that chymotrypsinogen has a molecular weight of about 25 000, a well-defined isoionic point, and is composed of a single polypeptide chain with a single half-cystine as the *N*-terminal residue and a single asparagine as the *C*-terminal residue.‡ The 25 000 molecular-weight unit contains 242 amino-acid residues.[12]

The relation of the primary structure of the enzyme to its catalytic function may be elucidated by a study

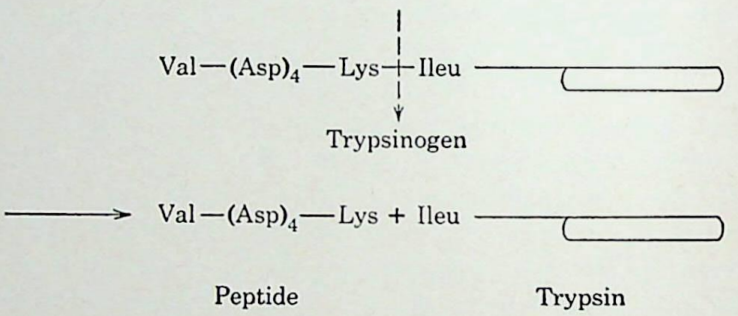


FIG. 1. Schematic representation of the tryptic activation of trypsinogen.

of reactions which involve the hydrolysis of a limited number of peptide bonds ("limited proteolysis"). Two types of reactions may be considered—namely, (1) those which involve the conversion of enzyme precursors to the active form (zymogen activation), and (2) those which lead to the partial degradation of active enzymes.

The controlled hydrolysis of one or more peptide bonds in trypsinogen or chymotrypsinogen causes these zymogens, having potential enzymatic activity, to become converted to active enzymes. As a result of this primary event, an intramolecular rearrangement occurs which presumably leads to the formation of the active site.[9] Figure 1 shows schematically the

TABLE III. Structural characteristics of chymotrypsinogen.

Molecular Weight	
Chemical analysis	25 081±336
Sedimentation, diffusion, *p*H 3.0	24 400
*p*H 7.5	24 000
Light-scattering	26 100±1000
X-ray	25 000±800
Osmotic Pressure	24 000±500
Isoionic point	9.4
N terminal	CySS/2
C terminal	None detectable in native protein
(Carboxypeptidase, hydrazinolysis)	

* In the case of one enzyme, ribonuclease, the entire sequence of amino acids had been recently determined by Moore, Stein, and co-workers at the Rockefeller Institute.[11] Ribonuclease is not a proteolytic enzyme, but catalyzes the hydrolysis of ribonucleic acid.

† The suffix *-ogen* denotes a precursor of an enzyme which, after certain enzymatic changes, becomes converted into an active enzyme. Such an enzyme precursor is also known as a zymogen.

‡ The *C*-terminal asparagine residue appears to be inaccessible to carboxypeptidase in intact chymotrypsinogen, and it remains inaccessible in the chymotrypsins. It can be detected after rupture of the disulfide bonds, by reduction or by oxidation.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

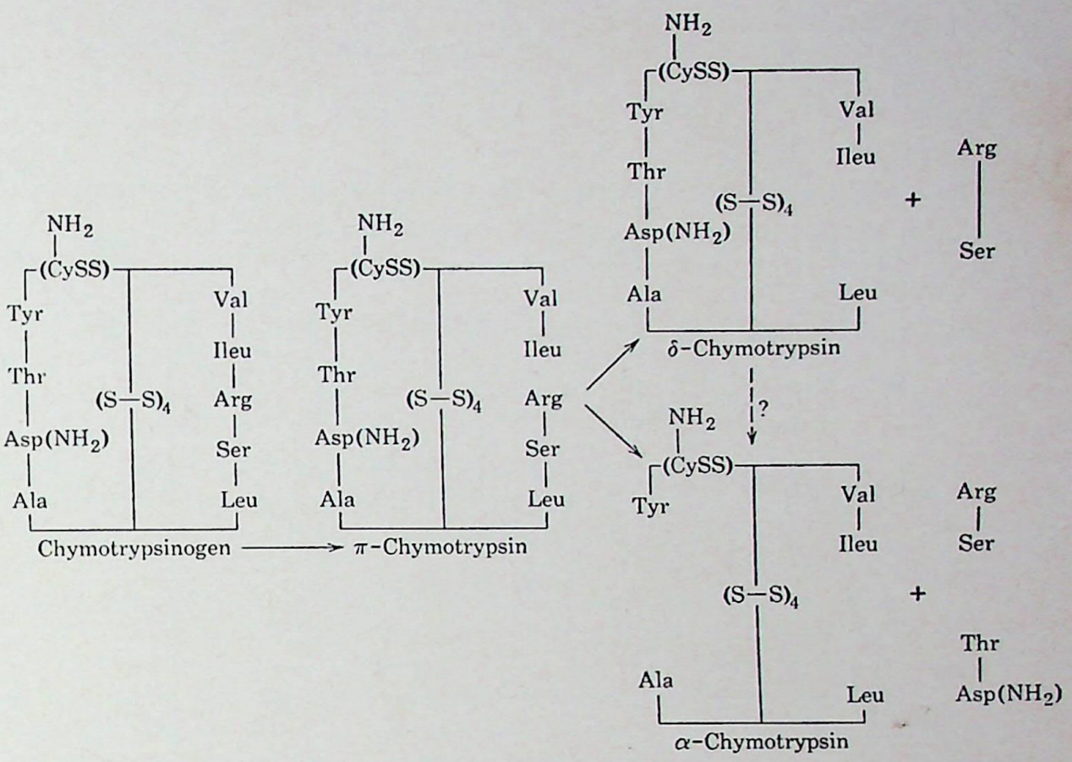


FIG. 2. Schematic representation of the activation of α-chymotrypsinogen. The first step, from left to right, is catalyzed by trypsin, all subsequent steps by chymotrypsin. The formation of δ- and α-chymotrypsins is believed to involve alternate rather than consecutive processes.

conversion of trypsinogen to trypsin. The trypsin-catalyzed hydrolysis of a specific, single peptide bond near the *N*-terminal region of the molecule, causes the hexapeptide, HVal—Asp—Asp—Asp—Asp—LysOH to be released and the *N*-terminal valine residue of the protein to be replaced by an isoleucine group. This seems to be the one and only chemical event which underlies the conversion of the inactive zymogen into the active trypsin enzyme.[13]

Figure 2 illustrates similarly the conversion of chymotrypsinogen to chymotrypsin. Here, only one of the seventeen bonds susceptible to hydrolysis by trypsin needs to be split in order to convert the zymogen into an active enzyme. These are fruitful processes in the sense that an active enzyme is formed from an inactive precursor.

Another example of peptide-bond hydrolysis relates to the partial degradation of an enzyme with or without concomitant loss of biological activity.[14] For instance, the removal from chymotrypsin of a few amino acids from the carboxyl end, with the aid of carboxypeptidase, yields a derivative which is still fully active. Similarly, in the case of ribonuclease, which consists of a single polypeptide chain, removal of as many as three amino acids from the carboxyl terminus yields an active derivative. If, however, four or more amino acids are removed from the same position, the product will be inactive. Another case is that of the ACTH, the adrenocorticotrophic hormone, which contains 39 amino acids. The removal of three amino acids from the *C*-terminal portion of the chain has no effect on hormonal activity, whereas the removal of one or two amino acids from the *N*-terminal portion, by use of leucine aminopeptidase, causes the hormone to become inactive. Perhaps the most dramatic case known today is that of papain, studied by Emil Smith and co-workers,[10] a proteolytic enzyme containing 180 amino acids. One hundred and twenty of these can be removed by stepwise hydrolysis with leucine aminopeptidase, and the derivative still maintains enzymatic activity, suggesting that the active center, including both the catalytic and specificity sites, is located in the carboxyl terminal region of the polypeptide chain and that the remainder of the molecule is not essential for over-all catalytic function.

In order to identify the chemical environment of the active site, it is necessary that it be labeled. In what follows, consider the reactions which are involved in the labeling of the active site of an enzyme and in the subsequent identification of the amino-acid composition of its environment.

The kinetics of the reaction of an enzyme—in this particular case, a proteolytic enzyme reacting with a substrate—indicate that at least two steps are involved. In the first step, the enzyme reacts with the substrate forming an acyl derivative; in the second step, this intermediate then reacts with water to form a free acid, and the enzyme is reconstituted.

$$\mathrm{RCOX+EH \rightarrow RCOE+HX},$$

$$\mathrm{RCOE+HOH \rightleftharpoons RCOOH+EH}.$$

It is this type of reaction which may be thought to underline the processes which Alberty describes for other systems (p. 177). In the reaction of a proteolytic enzyme with esters, the acyl-enzyme is formed first with the liberation of the alcohol, subsequent hydrolysis

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

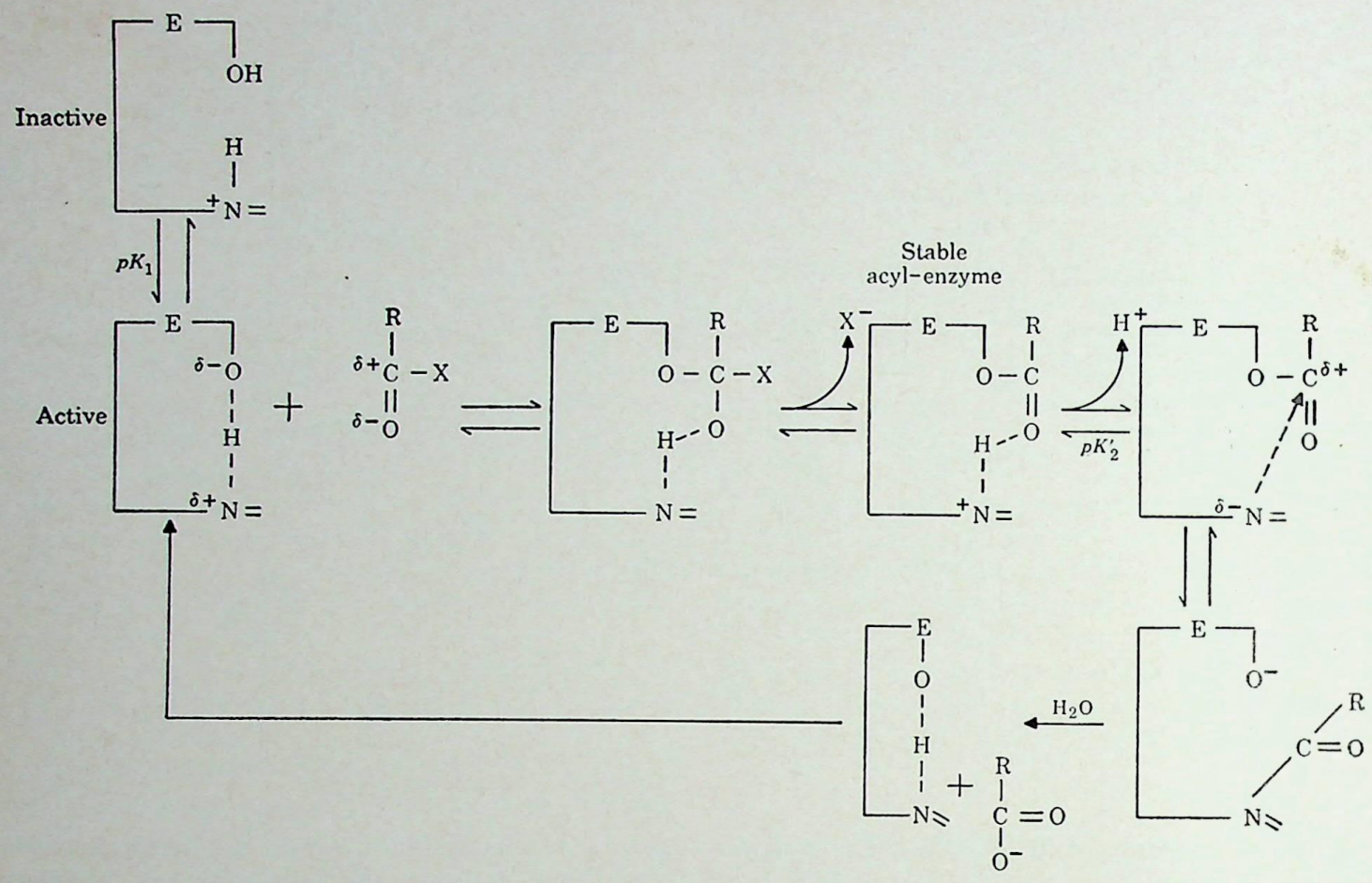


FIG. 3. Proposed mechanism of enzymatic ester hydrolysis, involving the interaction of the imidazole nitrogen of a histidine side chain and the hydroxyl group of a serine side chain.

of the acyl-enzyme leading to the free acid. This type of reaction has been studied in considerable detail, because substrates such as *p*-nitrophenylacetate, which contain a chromophoric group absorbing in the visible region of the spectrum, lend themselves excellently to precise kinetic studies.[15–17]

In the reaction of chymotrypsin or trypsin with organic phosphates such as diisopropyl phosphorofluoridate (DFP), which has been studied for a number of years, a phosphoryl-enzyme is formed instead of an acyl-enzyme. This is a convenient reaction for purposes of labeling, because the second step normally does not occur.

$$EH + F-\underset{\underset{O}{\|}}{P}(OC_3H_7)_2 \rightarrow E-\underset{\underset{O}{\|}}{P}(OC_3H_7)_2 + HF$$

(stable).

It requires the action of a more strongly nucleophilic reagent than water for the phosphoryl group to be removed from the enzyme. It is, therefore, possible to react an enzyme such as chymotrypsin with DFP^{32}, leading to the incorporation of the diisopropyl phosphate into the protein, and then to degrade the protein and see where the P^{32} went.

This type of study has shown that, in the case of trypsin, chymotrypsin, cholinesterase and in the case of the enzyme thrombin which is involved in the blood coagulation process, only one mole of DFP was specifically taken up per mole of enzyme, suggesting that these enzymes contain only one active center. In each case, the organic phosphate was found to be esterified to the hydroxyl group of serine. The numerous studies designed to identify the peptide sequence around the reactive serine residue have been described in considerable detail in recent reviews and are not considered here.[9,13,18] Suffice it to summarize, as shown in Tables IV and V, the composition and structure of the peptides obtained from DIP-trypsin and chymotrypsin, respectively. The serine group which is involved here must be

TABLE IV. DIP-peptides from esterases.

	α-Chymotrypsin	Trypsin	Liver aliesterase
DIP	1	1	1
Ser	1	1	1
Gly	2–3	2–3	2
Ala	...	...	1
Asp	1	1	...
Glu	...	...	1
Pro	1	1	1
Leu or Ileu	1	...	1
Val	...	1	...

TABLE V. Sequence of DIP-peptides.

α-Chymotrypsin: —gly—asp—ser(DIP)—gly—gly—pro—leu
Trypsin: —asp—ser—cyS—glu—gly—gly—asp—ser(DIP)—gly—pro—val—cyS—ser—gly—lys

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

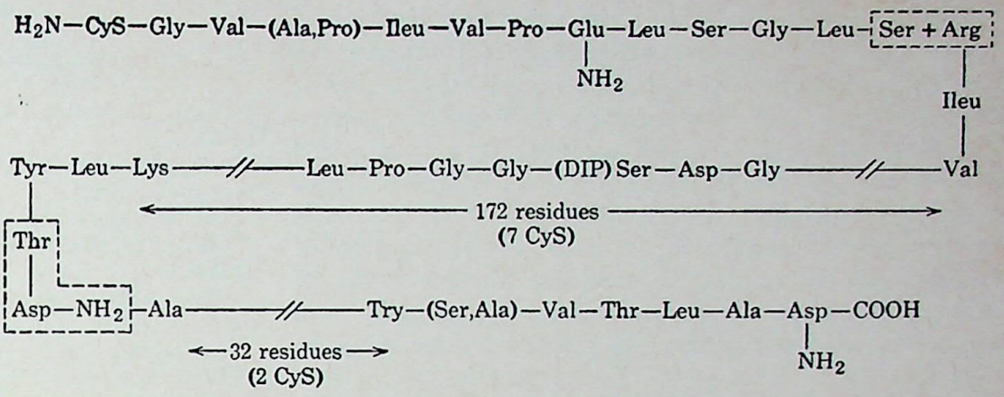


FIG. 4. Fragmentary representation of the amino-acid sequence of the single polypeptide chain of α-chymotrypsinogen. The dipeptide sequences delineated by the dotted lines are split off during the activation (see Fig. 2).

of unusual reactivity since there are twenty to thirty serine groups in trypsin and chymotrypsin, respectively, but only one of these is involved in the binding of DFP. Incidentally, the same seryl group is involved in the reaction of these enzymes with acylating agents such as *p*-nitrophenylacetate. Simpler seryl peptides do not behave in this way, and, if the enzyme is first denatured, the reaction with DFP does not take place at all. Conversely, if chymotrypsin is allowed to become acylated at *p*H 5, the acetyl group can be removed at this *p*H by reaction with hydroxylamine; but, if the protein is denatured by 8 molar urea after acylation, the acetyl group cannot be removed until denaturation is reversed, suggesting that separation and uncoiling of polypeptide chains removes in space another group or grouping which confers the high reactivity on this particular serine.[18] A number of lines of reasoning and experimental data has led to the conclusion that the second group involved in the enzymatic reaction is the imidazole group of a histidine side chain. This subject has been considered in detail in a recent review[13] and no attempt is made, therefore, to evaluate the experimental evidence for this hypothesis. The presence of histidine in the active region could account for the special reactivity of an adjacent serine residue, and, in a scheme developed by Cunningham, such an interaction has been proposed to include a hydrogen bonding between serine and histidine, as shown in Fig. 3. According to this scheme, the hydroxyl group of the serine and the nitrogen of the imidazole group of the histidine interact by hydrogen bonding, forming the active configuration; the *p*H dependence of the reaction is related to the *p*H dependence of the formation of these bonds.

The exact position of the serine and histidine within the structure of trypsin or chymotrypsin is unknown. In order to describe the structure of the active site, it would be necessary to know the complete amino-acid sequence, whereas, in fact, only partial sequences can be established, as shown in Fig. 4 for α-chymotrypsinogen. The exact tertiary structure of the polypeptide chain likewise needs to be elucidated. While this goal is still a long way off, nevertheless certain interpretations can be advanced on the basis of the data at hand, as shown schematically in Fig. 5, for the formation and the structure of the active site of trypsin.

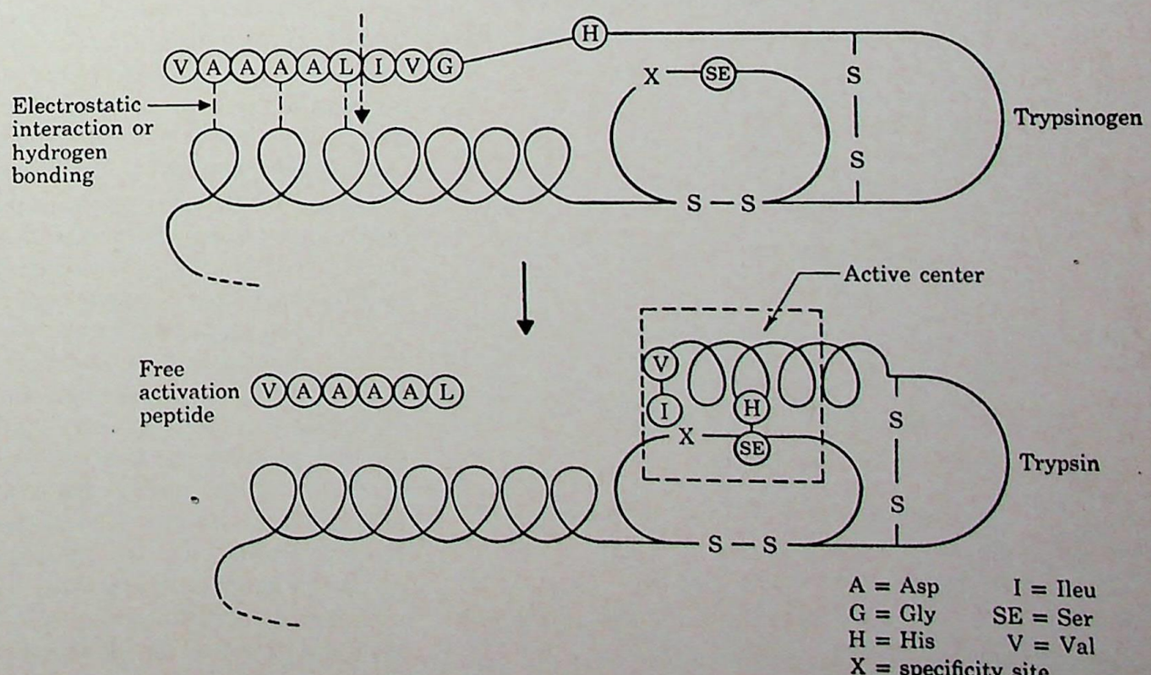


FIG. 5. Schematic representation of the structural changes involved in the tryptic activation of trypsinogen. Rupture of the lysyl-isoleucine bond in the *N*-terminal region (dotted arrow) leads to the liberation of the activation peptide, and causes the newly formed *N*-terminal region of the polypeptide chain to assume a more nearly helical configuration. This, in turn, permits a histidine and serine side chain to come into juxtaposition so as to form the esteratic site (see Fig. 3). The specificity side (X) is believed to be pre-existent in the zymogen molecule.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

The active site does not exist prior to activation of trypsinogen (or chymotrypsinogen), although the ability to bind the substrate may be pre-existent in the zymogen molecule. Furthermore, although the serine-histidine interaction would account for a generalized process of ester or peptide hydrolysis, it would fail to account for the high degree of specificity characteristic of proteolytic enzymes. It becomes of paramount importance, therefore, to elucidate those aspects of structure which differentiate between trypsin and chymotrypsin, and which, at the same time, may be correlated with their respective enzymatic specificities.

It is not fruitful to speculate about the reasons why enzymes are as large as they are just as no one has given a physical interpretation for the role of the complex and large structures of cholesterol or porphyrin. The fact that one has just begun to identify the functional relations among a few amino-acid residues in a molecule containing 200 or more does not justify the conclusion that the remainder is without function. On the other hand, the demonstration that *in vitro* one-half of a molecule may have similar if not identical functions as the whole does not lead to the compelling conclusion that, in a complex system such as living cell, the other half is of no consequence. It remains for the future to clarify fully what today can at best be considered as a promising beginning.

BIBLIOGRAPHY

[1] J. B. Sumner, J. Biol. Chem. **69**, 435 (1926).

[2] J. H. Northrop, M. Kunitz, and R. M. Herriott, *Crystalline Enzymes* (Columbia University Press, New York, 1948), second edition.

[3] M. Dixon and E. C. Webb, *Enzymes* (Academic Press, Inc., New York, 1958).

[4] M. Kunitz, J. Gen. Physiol. **32**, 241 (1948).

[5] M. A. Eisenberg and G. W. Schwert, J. Gen. Physiol. **34**, 583 (1951).

[6] B. L. Vallee, Advances in Protein Chem. **10**, 318 (1955).

[7] E. L. Smith, N. C. Davis, E. Adams, and D. H. Spackman in *The Mechanism of Enzyme Action*, W. D. McElroy and B. Glass, editors (The Johns Hopkins Press, Baltimore, Maryland, 1954), p. 291.

[8] B. L. Vallee and H. Neurath, J. Biol. Chem. **217**, 253 (1955).

[9] H. Neurath, Advances in Protein Chem. **12**, 320 (1957).

[10] J. R. Kimmel and E. L. Smith, Advances in Enzymol. **19**, 267 (1957).

[11] C. H. W. Hirs, W. H. Stein, and S. Moore in *Symposium on Protein Structure*, A. Neuberger, editor (Methuen and Company, London; John Wiley and Sons, Inc., New York, 1958), p. 211.

[12] P. E. Wilcox, E. Cohen, and W. Tan, J. Biol. Chem. **228**, 999 (1957).

[13] G. H. Dixon, H. Neurath, and J.-F. Pechère, Ann. Rev. Biochem. **27**, 489 (1958).

[14] C. B Anfinsen and R. R. Redfield, Advances in Protein Chem. **11**, 1 (1956).

[15] B. S. Hartley and B. A. Kilby, Biochem. J. **50**, 672 (1952).

[16] H. Gutfreund and J. M. Sturtevant, Biochem. J. **63**, 657 (1956).

[17] G. H. Dixon and H. Neurath, J. Biol. Chem. **225**, 1049 (1957).

[18] H. Neurath and G. H. Dixon, Federation Proc. **16**, 791 (1957).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

23
Molecular Structure of the Nucleic Acids

ALEXANDER RICH

Department of Biology, Massachusetts Institute of Technology, Cambridge 39, Massachusetts

ONE of the most useful, although simplified, views of the role of the nucleic acids in cells is that they function in a manner analogous to a punched tape in a computer. The punched tape and the nucleic acids are very long elements which direct the larger unit, either computer or cell, in which they are located. They both carry information through the linear arrangement of a few fundamental repeating units along their length. The nucleic acids are very long polymeric molecules which are built up by the repeated connection of a very few small molecules. It is generally assumed that the specificity of nucleic-acid function arises from the particular sequence of its constituent residues, as well as from their geometrical configuration.

There are two classes of nucleic acids, deoxyribonucleic acid (DNA) and ribonucleic acid (RNA), and they differ in structure as well as in function. However, they are made up of similar though not identical chemical units. The fundamental building block of the nucleic acids is the nucleotide. It is a complex molecule consisting of a purine or pyrimidine base, a sugar residue, and a phosphate group. In DNA, the sugar is deoxyribose, while in RNA, it is ribose. These sugars differ by the presence of a hydroxyl group on C_2'. Both of the nucleic acids have four types of bases, two purines and two pyrimidines, and three of these are found both in the deoxyribose and in the ribose polymers. DNA and RNA contain the purines adenine (A) and guanine (G), as well as the pyrimidine, cytosine (C) (see Figs. 3 and 4). In addition, RNA has the pyrimidine uracil (U), whereas DNA has the closely related pyrimidine, 5 methyl-uracil (thymine, T). Thus, both of the nucleic acids have a similar chemical composition, and only differ by the presence of a systematic hydroxyl group on each nucleotide of RNA and by the absence of a methyl group on one of the bases. While this description of the chemical composition of the nucleic acids is roughly correct, it should be pointed out that some nucleic acids have modified purine or pyrimidine residues, such as a $-CH_2OH$ group or a glucose residue attached to cytosine in the bacteriophages.

Both ribose and deoxyribose are five-carbon sugars which are in the furanose form—i.e., in the form of a ring involving four of the carbons and one oxygen. The nucleotides of DNA and RNA are connected by the same linkage through the phosphate group which is attached to the C_3' and the C_5' atoms of successive sugar residues. In a schematic way, the polynucleotide chain for both RNA and DNA can be written as shown in Fig. 1. It should be pointed out that the chain in Fig. 1 is asymmetric in that it has a direction which is most easily seen by the sense of the $C_3'-C_5'$ linkage in the sugar residue.

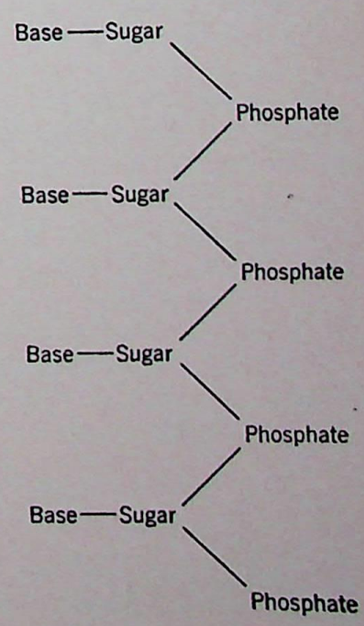


FIG. 1. Diagram of the nucleic-acid backbone. The bases are side chains on the repeating sugar-phosphate polymer chain.

INFORMATION TRANSFER AND THE NUCLEIC ACIDS

It is generally believed that DNA alone functions as the carrier of genetic information. This understanding is based upon the classic experiments of Avery who discovered bacterial transformation—that is, the ability of purified DNA from one bacterial species to alter the metabolic characteristics of another bacterial species in an inheritable manner. This interpretation of DNA function was further strengthened by the demonstration by Hershey and Chase[1] that DNA alone is the infective component and hence the carrier of genetic information in the bacterial viruses.

All cells of a given organism have the same DNA content. The only exception to this statement is to be found in spermatozoa and ova, where the DNA content is one-half of the normal amount. Further, if one analyzes the chemical composition of the DNA in all tissues of a given animal, it is found to be the same. Thus, it is believed that all cells contain the same set of DNA molecules. The DNA is located in the chromosomal material of the nucleus, and during cell division the DNA is replicated in some manner so that an equal amount is found in the two daughter cells, and with the same chemical composition as that found in the parental cell. One of the goals of molecular structural work in the nucleic acids is to discover the fundamental interpretation for these phenomena.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

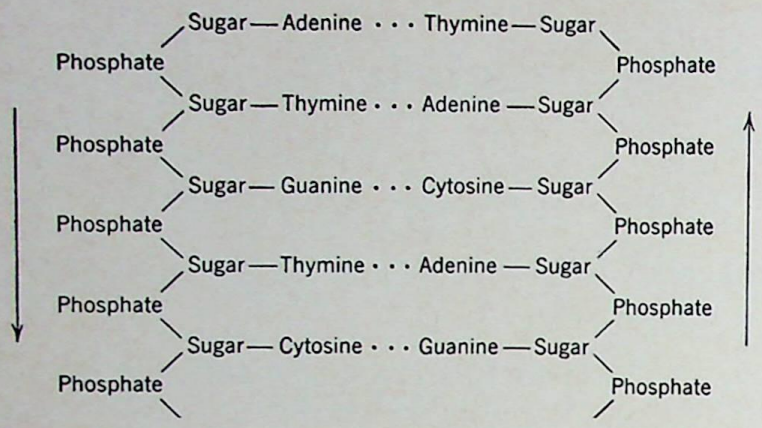


FIG. 2. Schematic diagram of DNA. The two chains are antiparallel, as shown by the arrows. The dotted lines between the bases represent hydrogen bonding. Although the chains are drawn as flat in the diagram, they are actually wound around each other in the molecule.

Although the amount of DNA is the same in each cell of a given organism, the amount varies from species to species. In general, the more complex species have more DNA. A bacterial cell has the order of 10^8 nucleotides in its DNA, which would make a molecular strand about 2 cm in length. In mammalian species, there are the order of 10^{10} nucleotides which corresponds to a total molecular length of 1 to 2 m/cell. Thus, the actual length of the primary coding material in a living cell is in the range of macroscopic dimensions and is much longer than the metabolic machine (or cell) which it directs. The same is often true for the punched tapes which are fed into computers.

MOLECULAR STRUCTURE OF DEOXYRIBONUCLEIC ACID

One of the most stimulating suggestions in molecular biology was a proposal made by Watson and Crick[2] that the molecular structure of DNA may consist of two polynucleotide chains helically wrapped around each other, with the sugar-phosphate chain on the outside and the purine and pyrimidine bases on the inside. They suggested that the purine and pyrimidine bases from the two chains are joined by hydrogen bonds to form specific pairs. Thus, the adenine residue hydrogen bonds with thymine, and guanine hydrogen bonds with cytosine. These hydrogen-bond pairs are specific in that only these combinations have the necessary stereochemistry to fit into the repeating lattice formed by the regular helical polynucleotide chains. In a schematic way, the DNA molecule is illustrated in Fig. 2 which shows the pairing relationship between the two polynucleotide chains. The arrows indicate the direction of the sugar-phosphate backbone. The two strands are organized in an antiparallel fashion so that the molecule looks the same even if it is turned about by 180°. If one ignores the varied base sequence, the backbone sugar-phosphate chains are organized about a diad axis perpendicular to the fiber axis and passing through the

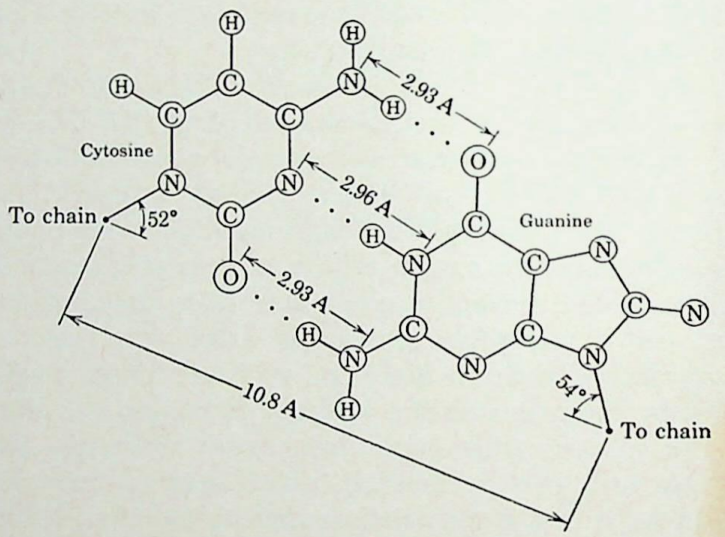


FIG. 4. Diagram showing the hydrogen bonding between guanine and cytosine in DNA. This pair is held together by three hydrogen bonds in contrast to the two found in the adenine-thymine pairing [from L. Pauling and R. B. Corey, Arch. Biochem. Biophys. **65**, 164 (1956)].

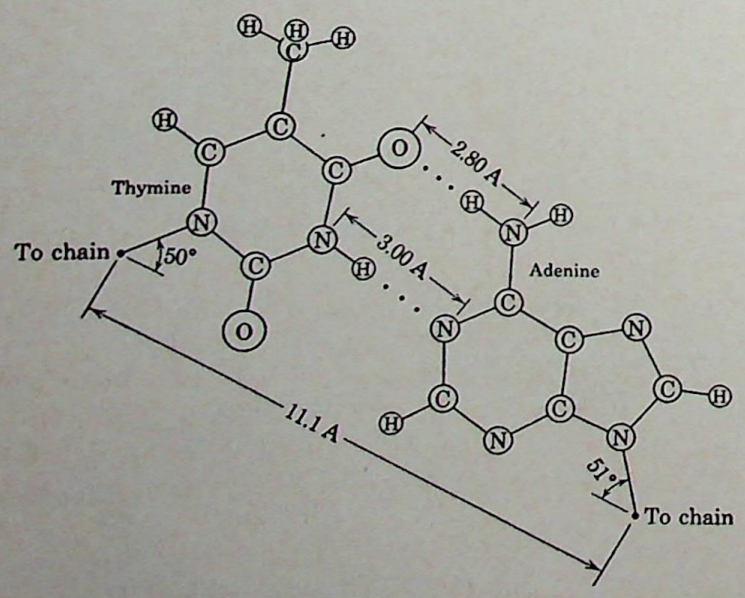


FIG. 3. Diagram showing the hydrogen bonding between adenine and thymine in DNA. The dimensions of the base pair are those discussed by Pauling and Corey.[3]

center of each base pair. The pairing of the bases is shown in Figs. 3 and 4. An important feature of the Watson-Crick hypothesis is the identity of the two types of base pairs. That is, the distance between the two sugar-phosphate chains must be the same both for the adenine-thymine pair and for the guanine-cytosine pair. In this way, both base pairs could fit into the helix interchangeably. Pauling and Corey[3] have made a critical survey of x-ray diffraction results obtained from crystals containing purines and pyrimidines. From this, they concluded that the cytosine and guanine residues are probably held together by three hydrogen bonds (Fig. 4), while the adenine-thymine residues are held together by two. The dimensions of the hydrogen-bond pairs suggested by Pauling and Corey are shown in Figs. 3 and 4. Within experimental error, the positions and angles of the two chains relative to the base pairs are identical.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

Early x-ray diffraction studies by Wilkins[4] and by Franklin,[5] and their collaborators showed that DNA is a helical molecule with 10 residues per turn. The gross features of the diffraction pattern were, at this early stage, shown to be compatible with the double-stranded form proposed by Watson and Crick. Wilkins and his collaborators have continued to carry out extensive studies on the diffraction patterns of DNA, and they are responsible for most of the knowledge concerning the detailed configuration of DNA.[6–8]

The DNA molecule exists in several forms. At lower relative humidities (about 70%), the molecule crystallizes in the *A*-form—that is, a face-centered monoclinic lattice with $a=22.2$ A, $b=40.0$ A, $c=28.1$ A, and $\beta=97.1$ A. This unit cell contains a repeat unit of two DNA molecules with the helical axis along c. The water content of this lattice is about 40%, and the bases in it are tilted about 25° from the fiber axis. In this form, the DNA is a true crystal and produces about 100 independent reflections. This implies that there is a high degree of regularity in the structure in all directions.

At higher relative humidities, the *B*-form which is paracrystalline appears with the molecules all parallel to each other but with random rotation about their molecular axes. The layer lines on these paracrystalline diffraction patterns show a continuous distribution of scattering intensity rather than the sharp spots characteristic of a crystalline lattice. In the *B*-form, the fiber-axis repeat is 34.6 A, and there is an extremely strong x-ray reflection on the meridian at 3.4 A. More recently, Wilkins and his collaborators[8] have been able to obtain truly crystalline diffraction patterns of the lithium salt of DNA in the *B*-configuration. In this form, the lithium DNA crystals are orthorhombic and have the dimensions $a=22.7$ A, $b=31.3$ A, and $c=33.6$ A. The axis of the helical molecules is along c, and two molecules pass through the unit cell.

The *B*-form of the DNA molecule is shown in Fig. 5 where atoms have been drawn with approximately their van der Waals radii. The base pairs are shown horizontally in the middle of the diagram, and the two sugar-phosphate chains are helically wrapped around the stacked bases. As can be seen in the diagram, there are two helical grooves which go round the DNA molecule. One of them is wider than the other because of the orientation of the sugar-base bonds shown in Figs. 3 and 4. The phosphates on the outside of the molecule are found at a radius of 9 A, and they are just over 7 A apart along a given chain. Wilkins and his co-workers have studied the organization of the polypeptide chain of protamine in a DNA-protamine combination.[7] The polypeptide chain is arranged as a third coaxial helix which fills the narrow groove in the DNA structure. In this position, the positively charged arginine side chains can interact with the negatively charged phosphate groups and stabilize the molecule.

Whenever a parent cell divides, the genetic information in that cell has to be replicated in order to insure continuity of inheritance to the daughter cell. The structural model of DNA suggested to Watson and Crick[9] a method for the replication of this molecule. They felt that the parent molecule could unwind as shown in Fig. 6 so that its two strands separated. These individual strands could then serve as a template for organizing the individual nucleotides which are necessary to make the second strand of the DNA daughter molecules. The template specificity is assured by the specificity of the hydrogen bonds between the purine-pyrimidine base pairs. In this way, a single molecule could twist about its axis, and two daughter helixes would form on the unraveled ends. Although this molecular model of genetic replication has not been established, recent experiments suggest that it is probably correct. This is a good example of the way in which a molecular structure can suggest a molecular mechanism.

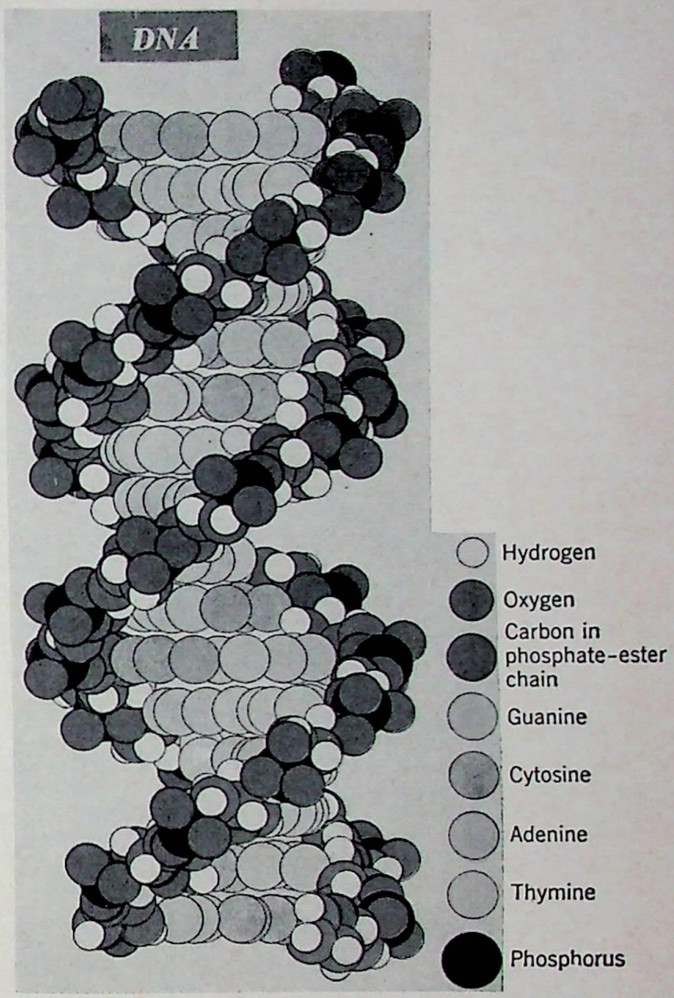


FIG. 5. A drawing of the DNA molecule using solid circles to illustrate atoms. It can be seen that there are two helical grooves of unequal size on the outside of the DNA molecule (after M. Feughelman *et al.*[7]).

ROLE OF RIBONUCLEIC ACID

In addition to carrying out its replication activities, it is necessary for the DNA molecule, acting as a genetic material, to influence and guide the metabolism of the

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

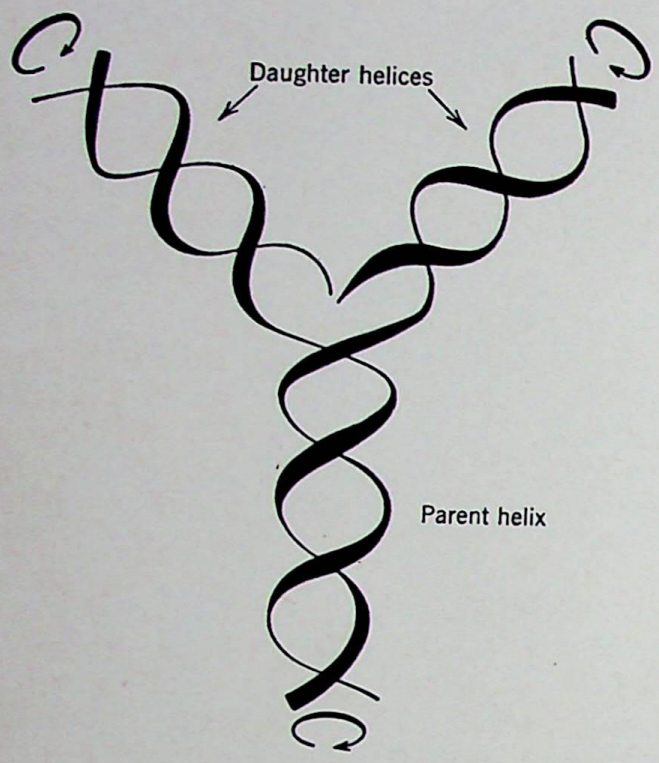


FIG. 6. Diagram showing a possible replication mechanism for DNA. The parent helix unwinds, and the two separated strands serve as sites for forming two daughter molecules. Both parent and daughter helices must wind simultaneously [from M. Delbrück and G. S. Stent in *The Chemical Basis of Heredity*, W. D. McElroy and B. Glass, editors (The Johns Hopkins Press, Baltimore, Maryland, 1957), p. 699].

cell. It is not known for certain how this is carried out. However, there is a large body of indirect information which suggests that this is carried out by using another nucleotide polymer which is somewhat similar to DNA—i.e., ribonucleic acid, which probably acts as an intermediary between DNA and the proteins which are synthesized and constitute the working chemical machinery of the cell.

The relationship between RNA and protein synthesis is not completely worked out at the present time. Proteins appear to be polymerized from their amino acids at the site of the small particulate bodies which are known as microsomal particles. These units are widely distributed in the cytoplasm, and appear as approximately spherical particles with a diameter of 180 A. They usually contain about half RNA and half protein, and it is likely that the RNA component plays a fundamental role in the protein-synthetic process, such as partly or wholly determining the sequence of amino acids. Thus, in order to control protein specificity, the information present in the DNA molecule in its nucleotide sequence may pass through the RNA molecule before ultimately emerging in a particular sequence of amino acids which define a particular protein. If these views are correct, one would like to know how it is that the DNA molecule "makes" RNA and how in turn this RNA molecule organizes amino acids. Unfortunately, at the present time, these questions cannot be answered.

The role of RNA in the metabolic cycle is not simple. In addition to being implicated in protein synthesis, it has also been demonstrated that RNA can function as a carrier of genetic information in a manner very similar to that of DNA. The pure RNA isolated from the tobacco mosaic virus is capable of infecting the tobacco leaf, and this infection ultimately produces a large number of new virus particles which carry the same genetic markers as those of the original virus.[10,11] Thus, RNA has some functions in common with DNA, but also appears to have unique activities as well.

Unfortunately, it is not possible to describe the molecular structure of RNA in a final manner, as is the case for DNA. RNA has been isolated in fibrous form and x-ray diffraction photographs taken of this material show a helical diffraction pattern with a fiber-axis repeat of approximately 26 to 28 A and strong meridional reflections at 3.3 and 4.0 A.[12] However, the diffraction photographs are not of sufficiently high resolution to allow one to make a unique structural interpretation. Experimental work on the structure of RNA offered little hope of achieving a solution to this molecular structural problem until the discovery of an enzyme capable of polymerizing synthetic polyribonucleotides.

Synthetic Polyribonucleotides

Grunberg-Manago and Ochoa[13] discovered an enzyme which converts nucleotide diphosphates into polyribonucleotides. The enzyme removes the terminal phosphate groups from the diphosphates and assembles the resultant nucleotide residues to form polyribonucleotides. Polymers obtained in this fashion resemble naturally occurring ribonucleic acid in that they have the same covalent ribophosphate backbone, and they have been shown to undergo similar enzymatic hydrolysis.[14] Polymers have been made from all of the purine and pyrimidine bases which are found in RNA. In addition, the enzyme will also make polymers which contain other purine or pyrimidine bases; e.g., polyinosinic acid has been made which contains the purine hypoxanthene, and recently a polymer has been made which contains thymine—i.e., a normal constituent of DNA but not of RNA. These polynucleotide molecules can be made either as pure molecules involving only one residue, or as copolymers involving two or more of the purine-pyrimidine side chains. The similarity between RNA and the synthetic polyribonucleotides can be shown by an x-ray diffraction study of synthetic copolymers, since they produce an x-ray diffraction pattern identical with that of native RNA.[15] This suggests that it might be possible to study the molecular configuration of the synthetic polyribonucleotides and thereby learn something about the configuration of naturally occurring RNA.

The synthetic polyribonucleotides are very reactive molecules. Soon after they were polymerized, it was shown that a complex formed when polyadenylic acid was mixed with polyuridylic acid.[16] Using x-ray diffraction analysis, it was found that these two mole-

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

cules wrap around each other in solution to form a two-stranded helical molecule very similar to naturally occurring DNA.[17] The discovery of this remarkable interaction has been followed by a variety of similar discoveries among the other polynucleotides. At present, we know of the existence of several of these elongate macromolecules which form two-stranded and three-stranded helical complexes.

Formation of Synthetic Two-Stranded Helical Molecules

If a dilute salt solution at neutral *p*H contains both polyuridylic acid (Poly U) and polyadenylic acid (Poly A), these two molecules complex together. The reaction is shown schematically in Fig. 7, where the bands represent the polynucleotide chains. On meeting each other, the molecules wrap about to form a two-stranded helix. Evidence for this reaction can be obtained from an x-ray diffraction study of a fiber drawn from a lyophilized mixture of the two polymeric species. The fiber has strong negative birefringence, and produces an x-ray diffraction pattern which has many similarities to a diffraction pattern of DNA. The distribution of scattering intensity is that which is characteristic of a helix: it has a large area on the meridian which is clear, and the scattering intensity is distributed in the form of a "cross" through the origin. The layer-line spacing varies slightly with humidity.[18] However, both DNA and (Poly A+Poly U) have a layer-line spacing of 34 A. This spacing represents the helical pitch of the molecule. From the strong meridional reflections in the region of 3 to 4 A, it can be shown that there are 10 residues per turn of the helix in both DNA and the (Poly-A+Poly-U) molecules. The birefringence of both materials is identical when the (Poly-A+Poly-U) molecules crystallize in a hexagonal lattice with a distance between the molecules of 28.8 A. This is approximately 6 A greater than that observed for the DNA molecule.

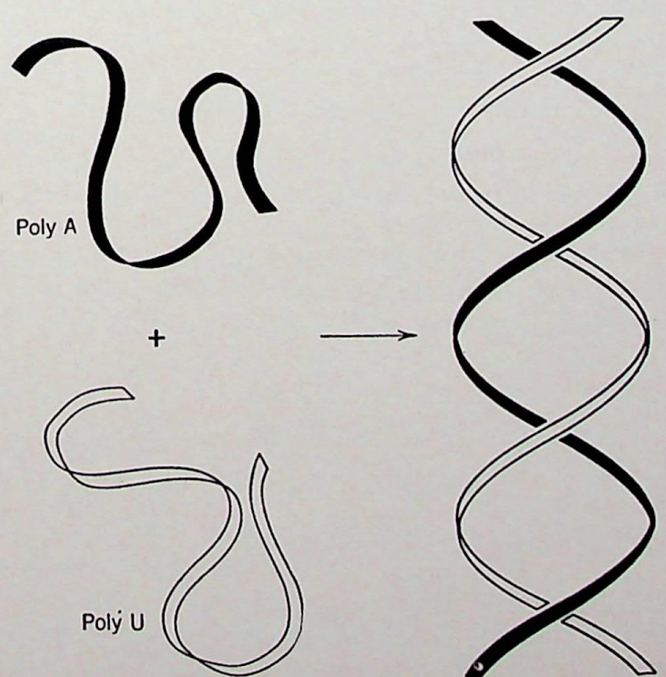


FIG. 7. Diagram illustrating the chemical reaction between the two polymers, polyadenylic acid (A) and polyuradylic acid (U). The irregular contours on the left represent the molecules in a random-coil configuration. After reacting on the right, the molecules are organized into a regular two-stranded helix. The bases connecting the two molecules are not shown in this diagram.

With the exception of the diameter of the molecule, the two diffraction patterns are similar enough to suggest that they arise from a similar helical structure. In the solution, the adenine residues of polyadenylic acid meet with and hydrogen bond onto the uracil residues of polyuridylic acid in a way which is identical to the kind of hydrogen bonding which occurs between adenine and thymine in DNA (Fig. 3). The only difference is that the uracil does not have the methyl group which is found on thymine. However, this does not affect the hydrogen bonding. Since the remainder of the molecule is similar to DNA, it forms the stablest structure possible—i.e., a DNA-like configuration. There is an additional hydroxyl group in the sugar residue of the polyribonucleotides relative to DNA, and this increases the diameter of the molecule slightly through its interaction with the other atoms of the sugar ring. This alters the hexagonal spacing mentioned above.

There are other methods which can be used to study the interaction between these two molecular species. When they react in solution, the optical density at 259 mμ decreases. This effect has been utilized in a quantitative study of the interaction.[19] A series of mixtures of polyadenylic acid and polyuridylic acid is made wherein the total concentration of phosphate groups remains constant, but the mole ratio of the two species varies continuously. The optical density is measured for this continuous series of solutions and the results are shown in Fig. 8. The dashed line shows the optical density of various mixtures of polyadenylic acid and polyuridylic acid at neutral *p*H in 0.1 molar sodium chloride, plotted as a function of mole ratio. It can be seen that the optical density falls quite sharply, and a minimum is reached at 50% mole ratio when the number of adenine residues in the solution just equals the number of uracil residues. This strongly suggests that a new species is being formed which is a 1:1 mixture of the two polymeric molecules. This interpretation is reinforced by studying this reaction in an ultracentrifuge, since there is an increase in molecular weight and sedimentation velocity when the two molecules combine.

Making careful measurements of the type shown in Fig. 8 for the 1:1 complex, it can be shown that over 95% of the residues have reacted, as judged by the sharpness of the drop in optical density. This is a measure of the high equilibrium constant for the reaction. One of the consequences of this figure is the inference that the reaction must be reversible in order to have all of these residues react.[20]

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

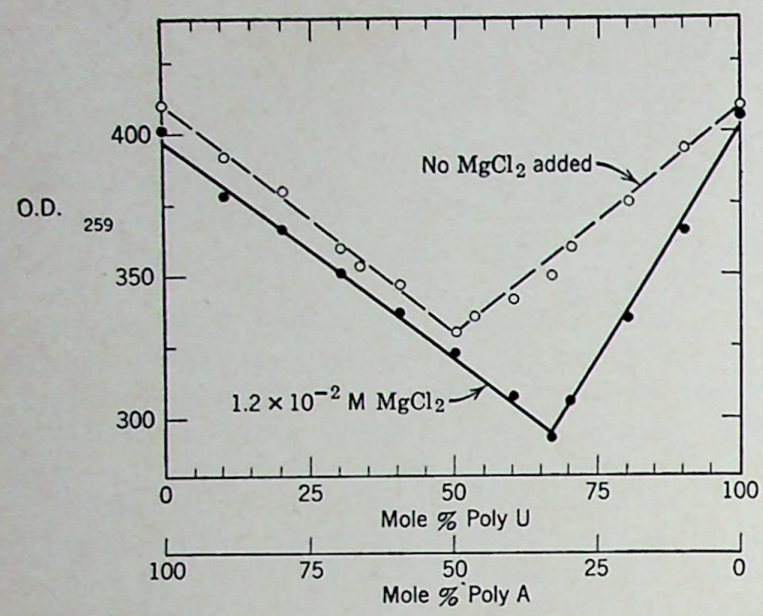


FIG. 8. The optical density of various mixtures of polyadenylic acid (A) and polyuradylic acid (U) at 259 mμ. The total number of moles of polymer is constant for all points, but the ratio of molecular species varies as indicated. All solutions are in 0.1 molar sodium chloride at neutral *p*H. The dashed line shows the formation of a 1:1 complex. The addition of a small number of divalent cations induces the formation of the three-stranded molecule [from G. Felsenfeld and A. Rich, Biochim. et Biophys. Acta **26**, 457 (1957)].

These experiments with mixtures of polyadenylic acid and polyuridylic acid clearly demonstrate the stability of the DNA configuration. In addition, they show that it is possible for the RNA covalent backbone to assume the form of a two-stranded complimentary duplex of the DNA type. This is significant because the work mentioned above on tobacco mosaic virus showed that the RNA molecule from the virus carries the genetic information residing in the virus. The molecule is also probably capable of carrying out the molecular replication necessary to virus multiplication in the leaf. In view of the fact that the RNA backbone can assume the DNA configuration, it seems quite reasonable to assume that the molecular replication of RNA may be carried out by a mechanism very similar to that involved in the molecular replication of DNA.

The reaction between polyadenylic acid and polyuridylic acid was perhaps not completely unexpected in view of the fact that DNA is composed of two strands held together by hydrogen-bonded purine-pyrimidine base pairs. Since uracil is so close to thymine, it is expected that the stability of (Poly A+Poly U) may be related to the stability of DNA itself. However, it is perhaps unexpected to find that it is possible to make a stable two-stranded helical molecule composed of polyadenylic acid and polyinosinic acid—i.e., a molecule very similar to the DNA molecule, except that it has purine-purine base pairs instead of purine-pyrimidine base pairs.

The evidence for this combination parallels that mentioned in the foregoing for polyadenylic acid and polyuridylic acid. When polyadenylic acid is mixed with polyinosinic acid under appropriate conditions, a lowering of the optical density occurs producing a minimum at the 1:1 mole ratio point, just as shown in Fig. 8.[21] Further evidence for the formation of this complex is also seen in ultracentrifuge experiments, since the complex has a larger molecular weight and sedimentation constant than either original molecule.

An x-ray diffraction pattern of a fiber of polyadenylic acid plus polyinosinic acid (Poly A+Poly I) is similar also to the *B*-form of deoxyribonucleic acid. The (Poly-A+Poly-I) molecules crystallize in a hexagonal array on the equator with $a = 24.4$ A. The fundamental screw operation for generating the (Poly-A+Poly-I) helix is a translation of 3.4 A and a rotation of 31.5°, just slightly less than the DNA and the (Poly-A+Poly-U) molecules.

The purine base in polyinosinic acid is hypoxanthine. It is closely related to guanine in that it has an oxygen on C_6 of the purine ring, even though it lacks the amino group present in guanine on C_2. It is likely that the hypoxanthine is in the keto tautomeric form and that it hydrogen bonds to the adenine residue as shown in Fig. 9. The keto oxygen of hypoxanthine is hydrogen bonded to the amino group of adenine, while the hydrogen on N_3 of hypoxanthine is bonded to the corresponding ring nitrogen in adenine. As can be seen when comparing this with Figs. 3 and 4, the hydrogen-bonding system has some similarities to what is observed in DNA. The major difference is the additional imidazole ring present in the hypoxanthine base. The hydrogen bonding shown in Fig. 9 could be used in the naturally occurring nucleic acids if the hypoxanthine base were replaced by guanine, since the additional amino group attached to C_2 of the purine ring would not introduce any steric interference.

Three-Stranded Helical Molecules

It was found that small amounts of divalent salts had a profound effect on the optical density-composition diagram in the Poly-A and Poly-U system.[22] The solid line in Fig. 8 shows the change brought about by making the solution 10^{-2} molar in magnesium chloride. A new minimum appears at 67% polyuridylic acid and 33%

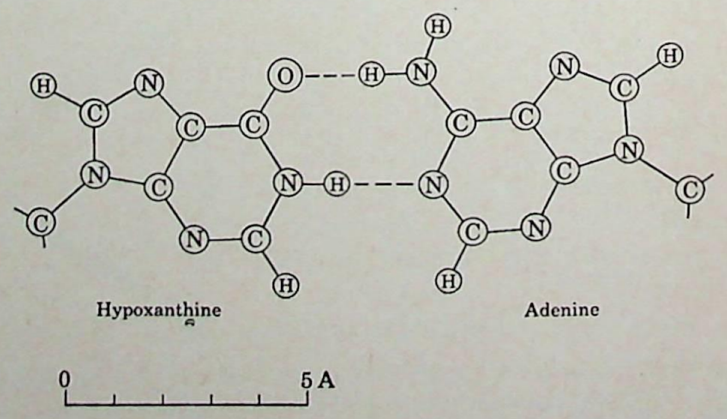


FIG. 9. Diagram showing the hydrogen bonding between the adenine base of polyadenylic acid and the hypoxanthine base of polyinosinic acid [from A. Rich, Nature **181**, 521 (1958)].

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

polyadenylic acid. The new minimum is quite sharp, indicating that the new complex is quite stable. This was the first indication that a three-stranded complex was forming from two strands of polyuridylic acid and one of polyadenylic acid.

Additional evidence for the formation of the three-stranded molecule is obtained by studying the sedimentation constant of the 1:1 complex as compared with the 2:1. The mean sedimentation constant for the three-stranded complex is about 50% greater than for the two-stranded complex. This increase would be expected if the two-stranded molecule were to take on a third strand. The third strand probably fills the deep helical groove in the two-stranded molecule which is similar to the deep groove seen in DNA. Since it would displace water molecules from that site and would not appreciably alter the frictional forces or shape factor of the molecule, the net density increment of the molecule over the solvent would result in approximately a 50% increment in sedimenting velocity.

It has been suggested that the second uracil residue is hydrogen bonding to the original adenine-uracil pair by forming two strong hydrogen bonds onto N_7 and N_{10} of the adenine ring. Such an additional third strand would not involve an increase in radius or helical pitch of the molecule, but could account for the approximately 50% increase in sedimentation velocity. X-ray diffraction photographs have also been obtained from this three-stranded complex.

The kinetics of the formation of the three-stranded molecules from a mixture of two-stranded (Poly-A + Poly-U) molecules and single-stranded Poly-U molecules have been investigated by measuring the optical density at 259 mμ as a function of time. Study of these curves for various concentrations of magnesium chloride or manganese chloride has shown that the reaction is second order for divalent cations; i.e., two divalent cations are present for each triplet of bases.

It is important to note that divalent cations do not have a unique role in forming the three-stranded molecule, since this complex will fully form in sodium-chloride solutions which are 0.7 molar. Thus, it is likely that the cations are necessary to overcome the electrostatic repulsion between the negatively-charged phosphate groups in the three polynucleotide chains. Divalent cations are much more effective than monovalent cations, probably because they form stable complexes with phosphate groups. Nonetheless, they are not necessary, since monovalent cations alone are capable of carrying out this reaction.

In a completely analogous fashion, the two-stranded polyadenylic-acid plus polyinosinic-acid molecule will take on a third strand of polyinosinic acid to become three-stranded.[21] This has been shown both spectrophotometrically as well as in the ultracentrifuge.

Up to this point, the discussion has been concerned with two- and three-stranded molecules composed of

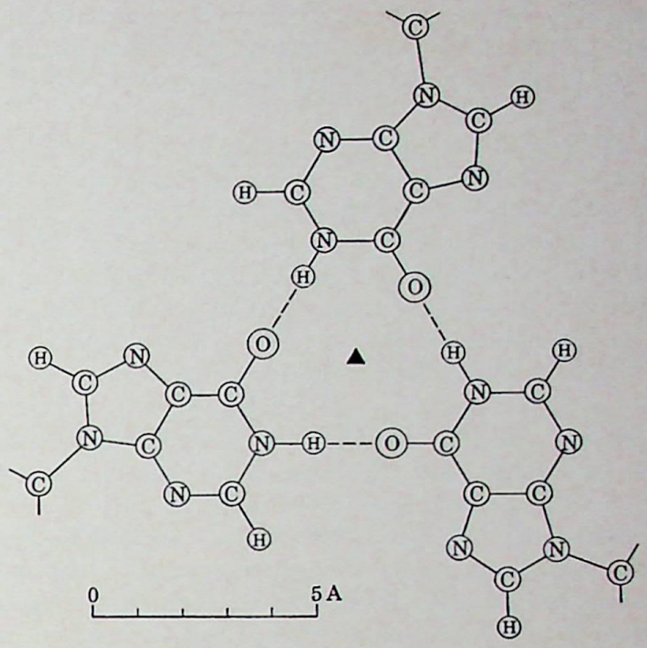


FIG. 10. Diagram showing the hydrogen bonding between three hypoxanthine bases in the three-stranded model of polyinosinic acid. The molecules are organized around a threefold rotation axis [from A. Rich, Biochim. et Biophys. Acta **29**, 502 (1958)].

different kinds of residues hydrogen bonded together. However, polyinosinic acid forms another kind of helical structure which involves only one type of molecule.[23]

If polyinosinic acid is prepared as a high molecular-weight polymer, it can be drawn into an oriented fiber which is negatively birefringent and which produces an unusual diffraction pattern when compared with the diffraction photograph of the mixtures described above. One unusual feature is that the first layer line is found at a spacing of 9.8 A, in contrast to the 30- to 40-A spacings discussed in the foregoing. Even more unusual is the appearance of the second layer line at 5.2 A and of the third intense meridional layer line at 3.4 A. These are nonintegral; i.e., they are not successive orders of one fundamental repeat distance. This feature and the fact that the first and second layer lines do not appear on the meridian of the diagram point uniquely to a helical configuration for the molecule. The meridional reflection at 3.4 A is undoubtedly attributable to the stacking of the purine residues at right angles to the fiber axis in agreement with the negative birefringence. The largest equatorial reflections occur at a spacing of 23.8 A.

This is an example of a helical diffraction pattern which points to a multiple-stranded structure, largely because it will not fit any single-stranded model. A three-stranded model will fit the diffraction data if all three strands are parallel to each other. The bases in this model are organized around a threefold rotation axis, utilizing the hydrogen-bonding system shown in Fig. 10. Here, the three hypoxanthine residues are hydrogen bonded in a cyclic manner, so that the keto oxygen attached to position 6 in the purine ring is hydrogen

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

bonded to the nitrogen at position 1 on an adjoining ring. The threefold rotation axis has the effect of decreasing the repeat distance along the fiber axis from 29.4 to 9.8 A.

Polyinosinic acid in solution can be converted from an organized helix into a random coil, and vice versa. This alteration is characteristic of all the polynucleotide molecular complexes. There are certain conditions under which they form stable, organized aggregates, and other conditions under which they separate in solution and are no longer organized. It is only when polyinosinic acid is in a single-chain random coil that it is able to react with polyadenylic acid to form the two- and three-stranded molecules already described; if polyadenylic acid and polyinosinic acid are mixed together in high salt concentration, no reaction occurs at all.

Several other combinations of polynucleotides have been studied and are mentioned only. Polyadenylic-acid chains combine with themselves to form two-stranded helixes in which both chains are parallel to each other. Polyinosinic acid, in addition to combining with itself and with polyadenylic acid, will also combine with polycytidylic acid.[24] Recently, it has been possible to synthesize polyribothymidylic acid.[25] This is a polymer with an RNA backbone even though it contains a DNA base. This molecule also combines with polyadenylic acid to form two- and three-stranded molecules, in a manner analogous to the combination with polyuridylic acid.[26]

DISCUSSION

The synthetic polyribonucleotides are extraordinarily reactive, and they can readily form a variety of multi-stranded helical structures, in some cases with molecules which are all alike, and in other cases with molecules which are different. There are, however, some generalities which emerge from this study which undoubtedly reflect some of the fundamental stabilizing features found in helical nucleic-acid molecules. For example, all of the helical molecules discovered so far have been either two- or three-stranded, but in no case has there been a stable single-stranded molecule. In the structures described in the foregoing, including DNA, the over-all architecture has been similar in that the purine or pyrimidine bases appear to be largely on the inside of the molecule, whereas the charged sugar-phosphate chain is on the outside of the molecule. The bases are hydrogen bonded together, usually with two or occasionally three strong hydrogen bonds holding each base. The molecules are stabilized by the van der Waals packing of the stacked planar purines or pyrimidines, in addition to the hydrogen bonding. It may be that a single-chain helix cannot form in a polyelectrolyte molecule such as the nucleic acids, because of the electrostatic repulsion. When two or more chains are present, the electrostatic repulsion in the helical molecule usually tends to pull the two chains closer to each other since they are coiled. In a single-stranded molecule, the effect of electrostatic repulsion would be only to elongate the molecule and break up any organized helical structure.

All of the synthetic organized helices are formed reversibly, and they can be made to assume a random coil in solution when conditions are appropriately modified. In most cases, a simple reduction of ionic strength is enough to drive the two chains apart; in other cases, altering the *p*H is sufficient. Usually, there is only a given *p*H range over which a multistranded molecule is stable. For example, the polyinosinic-acid helix breaks up when the *p*H is raised above 10 owing to the fact that a proton necessary in the hydrogen-bonding system is removed. Similarly, the other helices are stable only over *p*H ranges in which the necessary tautomeric forms exist. Studies such as these are useful in developing an understanding of the stability of the naturally occurring nucleic acids and, since several of these molecules are similar to DNA, they can be used as model systems for studying nucleic-acid reactivity.

In this regard, an especially attractive hypothesis may be made concerning the three-stranded molecules. It is possible that these may be analogs for the formation of a single-stranded RNA by a two-stranded DNA. DNA has a deep helical groove in it between the two strands. It is this groove which is filled in the (polyadenylic plus polyuridylic) molecule by the oncoming strand of polyuridylic acid, and, as such, it may be an example of a physiologically important type of reaction. Since DNA itself has two kinds of base pairs, there are a total of four different sites on the DNA molecule. These four types of sites may serve as templates for the four kinds of ribonucleotides which must be polymerized together to make the RNA molecule. Schemes such as this have been worked on by several investigators for many years, since it is an attractive and simple method for transferring sequence information from DNA to RNA. However, there has been as yet no convincing demonstration of a detailed molecular mechanism which has the requisite specificity. Further work is necessary to fully evaluate mechanisms of this sort.

Despite the variety of polynucleotide structures which are now understood, the structure of naturally occurring RNA itself remains unknown. There is undoubtedly a difference in the configuration of the RNA which is found in the microsomal particle from that found in the small RNA molecules present in the soluble supernatant. In addition, it is possible that the nuclear RNA is in yet a different configuration. The most interesting configuration for RNA is perhaps that which is found in the microsomal particle. According to the current hypothesis, the nucleotide sequence in this molecule is translated by some means into the amino-acid sequence in the proteins which are synthesized at that site. The configuration of RNA in these particles is indeed an interesting problem, and there is no way of

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

knowing whether or not the studies on the synthetic polyribonucleotides will yield an answer which is at all applicable to the problem. Nonetheless, these studies are producing a variety of structures and configurations and it can only be hoped that future work will yield promising results in this most interesting aspect of the problem.

BIBLIOGRAPHY

[1] A. D. Hershey and M. Chase, J. Gen. Physiol. **36**, 39 (1952).
[2] J. D. Watson and F. H. C. Crick, Nature **171**, 737 (1953).
[3] L. Pauling and R. B. Corey, Arch. Biochem. Biophys. **65**, 164 (1956).
[4] M. H. F. Wilkins, A. R. Stokes, and H. R. Wilson, Nature **171**, 738 (1953).
[5] R. E. Franklin and R. G. Gosling, Nature **171**, 740 (1953).
[6] M. H. F. Wilkins, W. E. Seeds, A. R. Stokes, and H. R. Wilson, Nature **172**, 759 (1953).
[7] M. Feughelman, R. Langridge, W. E. Seeds, A. R. Stokes, H. R. Wilson, C. W. Hooper, M. H. F. Wilkins, R. K. Barclay, and L. D. Hamilton, Nature **175**, 834 (1955).
[8] R. Langridge, W. E. Seeds, H. R. Wilson, C. W. Hooper, M. H. F. Wilkins, and L. D. Hamilton, J. Biophys. Biochem. Cytol. **3**, 767 (1957).
[9] J. D. Watson and F. H. C. Crick, Nature **171**, 964 (1953).
[10] A. Gierer and G. Schramm, Nature **177**, 702 (1956).
[11] H. Fraenkel-Conrat, J. Am. Chem. Soc. **78**, 882 (1956).
[12] A. Rich and J. D. Watson, Proc. Natl. Acad. Sci. U. S. **40**, 759 (1954).
[13] M. Grunberg-Manago and S. Ochoa, J. Am. Chem. Soc. **77**, 3165 (1955).
[14] L. A. Heppel, P. J. Ortiz, and S. Ochoa, J. Biol. Chem. **229**, 679 (1957).
[15] A. Rich, Ann. N. Y. Acad. Sci. Special Publ. **5**, 186 (1957).
[16] R. C. Warner, Federation Proc. **15**, 379 (1956).
[17] A. Rich and D. R. Davies, J. Am. Chem. Soc. **78**, 3548 (1956).
[18] A. Rich in *The Chemical Basis of Heredity*, W. D. McElroy and B. Glass, editors (The Johns Hopkins Press, Baltimore, Maryland, 1957), p. 557.
[19] G. Felsenfeld and A. Rich, Biochim. et Biophys. Acta **26**, 457 (1957).
[20] G. Felsenfeld, Biochim. et Biophys. Acta **29**, 133 (1958).
[21] A. Rich, Nature **181**, 521 (1958).
[22] G. Felsenfeld, D. R. Davies, and A. Rich, J. Am. Chem. Soc. **79**, 2023 (1957).
[23] A. Rich, Biochim. et Biophys. Acta **29**, 502 (1958).
[24] D. R. Davies and A. Rich, J. Am. Chem. Soc. **80**, 1003 (1958).
[25] B. E. Griffin, A. R. Todd, and A. Rich, Proc. Natl. Acad. Sci. U. S. **44**, 1123 (1958).
[26] Author's unpublished observations.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

24
Biosynthesis of Nucleic Acids

ARTHUR KORNBERG

Department of Microbiology, Washington University School of Medicine, St. Louis 10, Missouri

INTEREST in the biosynthesis of nucleic acids stems from the crucial importance of these compounds in heredity and reproduction and in the development of the cellular machinery.

Every cell has DNA in its nucleus and, from a variety of considerations, DNA is regarded as the prime source of genetic information. Consider a cell which just has been pinched from its mother and which has little more than a faithful replica of the maternal DNA. In its growth to maturity, the young cell will have to make the enzymatic machinery to carry on energy metabolism on an expanding scale. At the same time, the cell must synthesize all of the special structures (be they lipid, carbohydrate, protein, or of any other special nature) that will characterize the shape, structure, appearance, and behavior of the cell. The sequence can be represented, oversimplified, as

$$\text{DNA} \rightarrow \text{enzymes} \rightarrow \text{everything else.}$$

From all indications, it can be said with some confidence that the structure of DNA is fixed, that it does not undergo turnover under the most extreme variations of cellular nutrition and physiology, and, facetiously, that it is immutable, except for mutations. Aside from having to serve as an original template for the synthesis of the enzymes, it has one other major function and that is to provide the template for the synthesis of a replica of itself in the reproductive act of forming a daughter cell. Discussion of the biosynthesis of DNA is limited here to its replication during reproduction.

RNA BIOSYNTHESIS

What is known about the function and biosynthesis of RNA? In contrast with DNA, this problem appears more complex. RNA occurs both in the nucleus and in a variety of cytoplasmic structures. Its concentration in the cell may vary within wide limits depending upon the age, nutrition, and other environmental factors, and it obviously undergoes considerable turnover and replacement. While the DNA of the chromosome might be regarded as a single, exceedingly complex molecular unit, RNA is, by comparison, physically and metabolically heterogeneous and there is no reason to speak of RNA in the singular, except in the generic sense that one regards protein.

It would appear from current studies that RNA may provide a link between DNA and protein synthesis, translating the information in the DNA code to a form that is a proper template for the manufacture of distinctive proteins. What is known, in fact, from the work of Zamecnik *et al.*[1] with animal cells, is that RNA is essential for the assembly of amino acids into polypeptides, and from the work of Berg[2] with a bacterial system, is that only about 5 to 10% of the RNA seems to be active in the fixation of amino acids. This function of RNA as a link between DNA and enzyme synthesis is at present an educated and attractive speculation only.

Regarding the biosynthesis of RNA, the problem is complicated by its heterogeneity and by ignorance of its functions. There is a large and rapidly growing literature on the subject, but brief reference is made to only a few of the contributions.

(1) Ochoa[3] isolated from a variety of bacterial cells an enzyme which condenses various nucleoside diphosphates to high molecular-weight ribonucleic-acid-like polymers. These, discussed by Doty (p. 107) and by Rich (p. 191), are polymers of adenylic, cytidylic, or uridylic acid, or of other nucleotides or mixtures of them. As seen in Fig. 1, the reaction involves a condensation between the hydroxyl group in position 3 of one nucleotide with the inner phosphate group of the other nucleotide, with a consequent elimination of inorganic orthophosphate (Pi). There is little free-energy change in this reaction and, as indicated, the polymers are split by inorganic phosphate at physiological levels. For this enzyme to form the highly specific polymers required in protein synthesis, it must have controls in the cell from which it has escaped upon isolation. A reasonable possibility is that this enzyme provides for the conservation and storage of nucleotide units which can be called upon to form the adenylic, uridylic, cytidylic, and other coenzymes essential for energy metabolism, for carbohydrate, protein, and lipid biosynthesis, and for specific types of RNA.

(2) Several groups of investigators[4–9] have recognized, during the past year or so, the existence of enzyme systems which extend ribonucleic-acid chains by one or a few nucleotide units, and which employ nucleoside triphosphates rather than the diphosphates. There are suggestive indications of relationships of these reactions to amino-acid fixation leading to polypeptide formation.

(3) A reaction which has the intriguing possibility of relating RNA synthesis to DNA is the one recently described by Hurwitz.[10] An enzyme from *E. coli* was found which required DNA for the incorporation of ribonucleotides into a polymeric structure. With a mixture of radioactive ribonucleoside triphosphates and deoxynucleoside triphosphates as substrates, the prod-

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

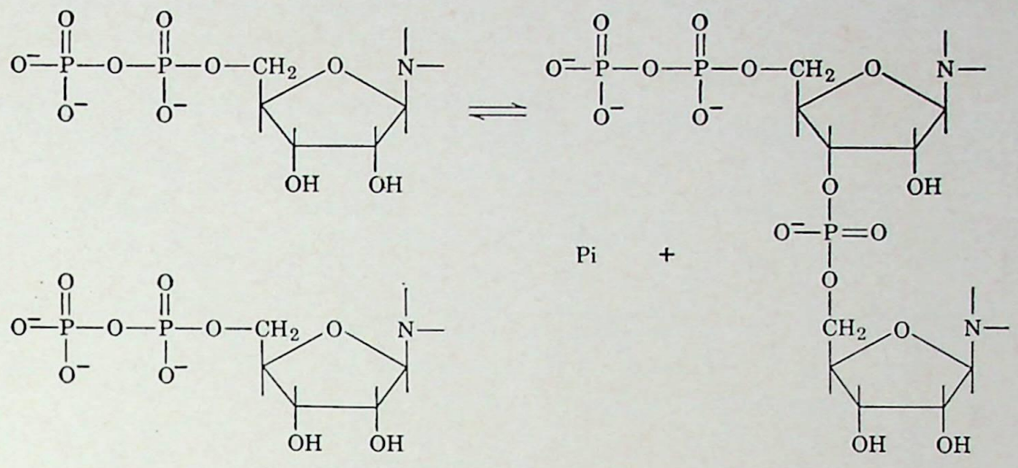


FIG. 1. Mechanism of RNA synthesis from ribonucleoside diphosphates.

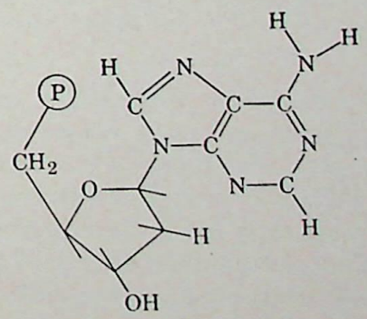
uct formed was shown to contain the incorporated units in typical 3′-5′-phosphodiester linkage. Samples of RNA from several sources failed to replace the DNA requirement. Further purification of this enzyme system should clarify the specificities of the reaction and they are awaited with great interest.

(4) Turning now to studies with whole cells which attempt to relate nucleic-acid and protein synthesis, mention should be made first that, by the addition of chloromycetin to growing bacterial cultures, it is possible to suppress protein synthesis completely and still maintain the new formation of RNA and DNA. The synthesis of RNA, however, can be shown to be dependent upon the availability of all of the amino acids essential for protein synthesis.[11–13] Thus, a strain of *E. coli* requiring tryptophan for growth fails to make RNA in the presence of chloromycetin unless tryptophan is added to the medium. Yet this and other amino acids for which this effect has been demonstrated do not appear to be precursors of RNA. A dependence of RNA synthesis upon amino acids, despite the exclusion of protein formation, is implied, therefore, but an enzymatic basis for these observations as yet is not apparent.

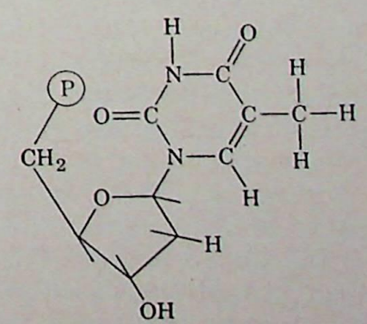
(5) Finally, the studies of RNA biosynthesis in virus-infected bacterial cells are noteworthy. In a logarithmically growing population of bacterial cells, there is an exponential increase of protein, RNA, and DNA. At the time of virus infection, as with *T2* infection of *E. coli*, synthesis of bacterial components ceases abruptly and the cellular machinery is devoted completely to the synthesis of viral protein and viral DNA. Since the virus lacks RNA, there is—or rather, there was thought to be—no RNA synthesis. Recent tracer studies by Volkin and Astrachan[14] have shown a burst of RNA synthesis, very small in amount but distinctive in composition. One interpretation of these findings is that this RNA synthesis is in response to the unique instructions carried in the viral DNA, and is a necessary prerequisite for the formation of the proteins essential to virus formation. There is here a highly simplified situation in which to study a form of RNA biosynthesis that is relevant to the link between DNA and protein and that is inviting to biochemical study.

DNA BIOSYNTHESIS

To review briefly the chemical structures which serve as the basic building units of the DNA molecule (Fig. 2): Deoxyadenylate (deoxyadenosine 5′-phosphate) is composed of a purine base linked by a glycosidic bond to the sugar that is unique to DNA, 2-deoxyribose. It is in the form of a furanose ring and lacks an oxygen at the 2-carbon position; esterified at carbon-5 is a phosphate residue. Thymidine 5′-phosphate, or thymidylate, is composed of a pyrimidine, thymine, linked as in the case of adenine to a 2-deoxyribose 5′-phosphate. The methyl group on the 5 carbon distinguishes thymine from uracil, also a pyrimidine, which is a component of



Deoxyadenosine 5′-P



Thymidine 5′-P

FIG. 2. Structures of a purine and a pyrimidine deoxynucleotide.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

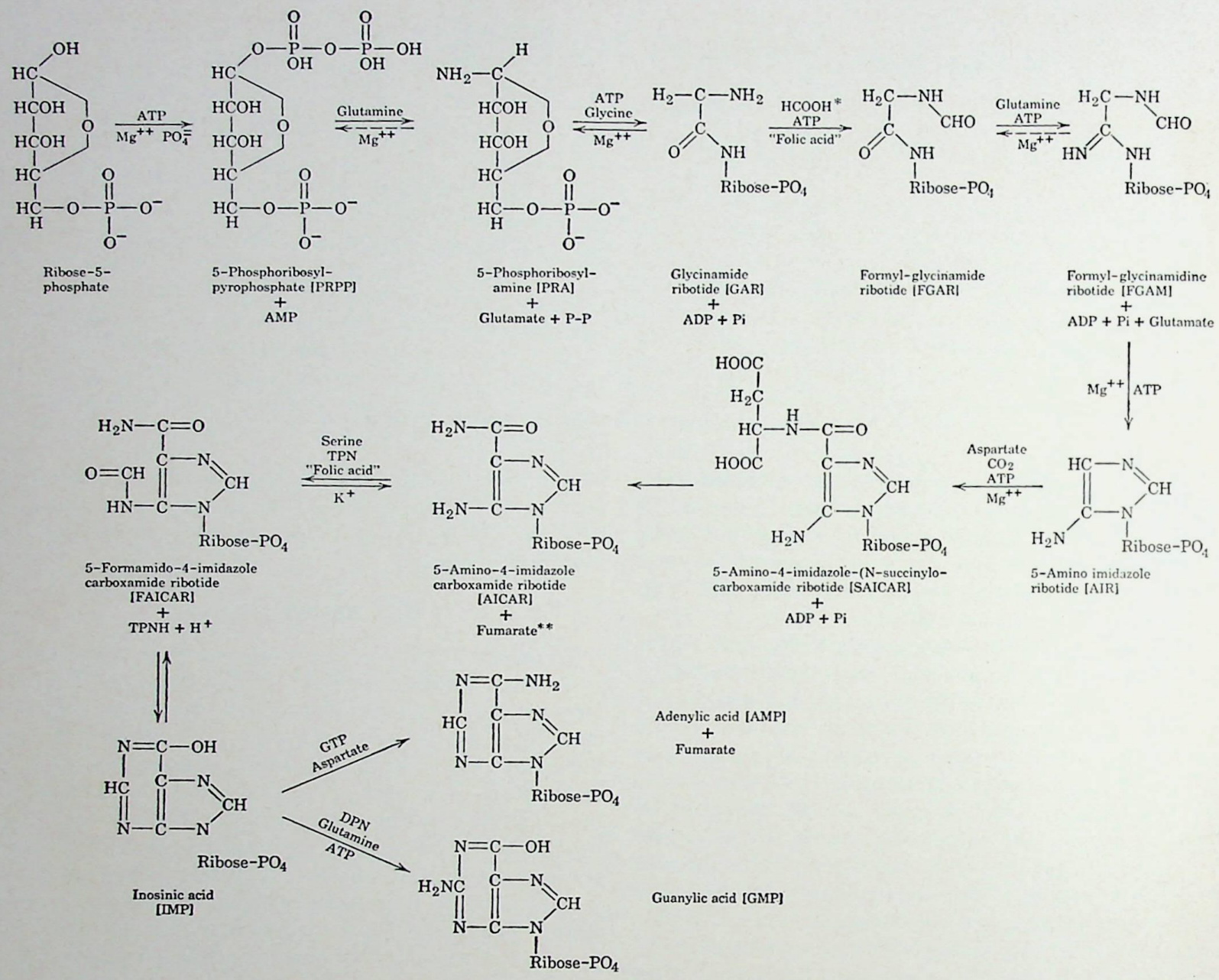


FIG. 3. Schema of enzymatic synthesis of purine ribonucleotides [from J. M. Buchanan, J. G. Flaks, S. C. Hartman, B. Levenberg, L. N. Lukens, and L. Warren, *Chemistry and Biology of Purines*, Ciba Foundation Symposium (J. and A. Churchill, Ltd., London, 1957), p. 233].

ribonucleic acid and of coenzymes prominent in carbohydrate metabolism. These and two other deoxynucleotides—a purine one, deoxyguanylate, and a pyrimidine one, deoxycytidylate—comprise the four deoxynucleotides which commonly occur in samples of DNA from bacterial, plant, and animal cells, and from some of the viruses. How are these units assembled?

This type of biosynthetic question has been approached in several ways. Earlier, Calvin (p. 147) and Roberts (p. 170) described the use of isotopic tracers to chart biosynthetic pathways. A suspected intermediate containing an isotopic marker is administered in the medium, and the cellular constituents which incorporate this marker are identified after brief or extended time exposures. A variation of this technique is to label the exclusive carbon or nitrogen source of the cell (e.g., C^{14} glucose or N^{15} ammonia), and then to determine whether or not certain (unlabeled) compounds in the medium can reduce the specific radioactivity of particular molecules in the cell under study.

Quite another approach involves the use of a mutant which accumulates intermediates because of an enzymatic deficiency at some point in the biosynthetic process. The facility with which mutants can be selected for service in a particular situation has made this technique extremely versatile and effective. These and other quite diverse methods have been used with considerable success in mapping pathways in the intact cell. But no combination of these methods can establish unequivocally a biosynthetic sequence! The walls and membranes of the cell form an "iron curtain" which either prevents certain molecules from entering the cell or alters them upon their entry or exit. It is a curtain which obscures the details to the point where a spur from a pathway is indistinguishable from the main line. Hence, a study of the enzymatic apparatus of the disrupted cell is an indispensable approach to the problem of biosynthesis. The objective of such an enzymatic attack on the problem is the isolation from cells of separate enzymes, each of which effects a single,

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

chemically rational step. These steps arranged in proper sequence should lead to the total synthesis of complex molecules. To assume real significance, the rates and conditions under which a reconstructed biosynthetic pathway operates ultimately must be reconciled with all of the isotopic, nutritional, and other observations made with the intact cells.

Turning now to what is known about the details of the enzymatic synthesis of the nucleotides (Fig. 3), one sees an outline of the discrete steps in the synthesis of adenosine 5′-phosphate starting with ribose 5-phosphate, glycine, glutamine, and, as an energy source, adenosine triphosphate (ATP). This knowledge is largely the result of the work of Buchanan[15] and of Greenberg[16] and their colleagues. A comparable scheme could be shown for the pyrimidine nucleotides. The origin of the deoxyribonucleotides would appear, from current indications, to derive from the ribonucleotides by a direct reduction step.

Several years ago, experimental work was begun by our laboratory to determine the biochemical mechanism for the replication of DNA chains. One might suggest many permutations for how such chains might be

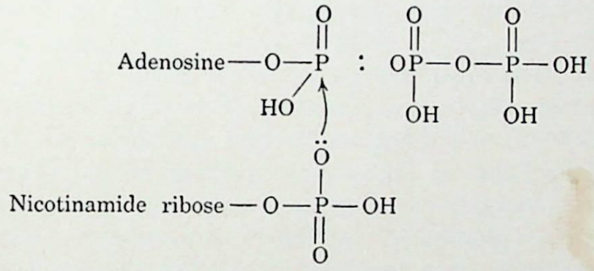


FIG. 4. Nucleophilic attack of nicotinamide mononucleotide on ATP in DPN synthesis.

assembled. We were guided by the knowledge of how the simplest and best known of the complex nucleotides, the coenzymes, are synthesized by the cell.[17–19] For example, the enzymatic synthesis of diphosphopyridine nucleotide (DPN) (Fig. 4) involves a condensation of ATP with nicotinamide mononucleotide and a resultant elimination of inorganic pyrophosphate (PP). Similarly, in the synthesis of flavin adenine dinucleotide (FAD), or coenzyme-*A*, there is a reaction of flavin mononucleotide or pantetheine phosphate, respectively, with ATP, and again inorganic pyrophosphate is eliminated (Fig. 5).[20] In each case, the adenyl coenzyme is produced by a reaction with an activated adenyl derivative such as ATP. More recently, the enzymatic synthesis of the adenyl derivatives of the fatty acids and amino acids have been shown to proceed by a similar mechanism.[21] What is shown in Fig. 5 to apply to the synthesis of adenyl derivatives applies also to the synthesis of uridyl, cytidyl, and guanosyl coenzymes.[22–24] Along lines proposed by Koshland,[25] one may visualize (Fig. 4) a nucleophilic attack on the activated adenyl derivative, in the case of DPN synthesis, by nicotinamide mononucleotide, to form an anhydride bridge with the adenyl group and to displace inorganic pyrophosphate. By analogy, it can be conjectured that chains of nucleotides might be formed by a reaction of the end of a chain with an activated nucleotidyl molecule, such as a nucleoside triphosphate. In other words, it is assumed that the basic building block is a 5′-phosphate ester of a deoxynucleoside, activated perhaps by linkage to a pyrophosphate group. As shown in Fig. 6, the end of a DNA chain might be

1. ARPPP(ATP) + NRP(nicotinamide mononucleotide) $\rightleftharpoons$ ARPPRN(DPN) + PP

2. ARPPP + FRP(riboflavin-P) $\rightleftharpoons$ ARPPRF(FAD) + PP

3. ARPPP + Pantetheine-P $\rightleftharpoons$ ARPP-Pantetheine + PP
   ATP → ARPP-Pantetheine (CoA)
   P

4. ARPPP + $RCH_2COOH \rightleftharpoons ARPOCCH_2R$ + PP
   ‖ O
   or or
   $RCHNH_2COOH$ $ARPOCCHNH_2R$ + PP
   ‖ O

FIG. 5. Synthesis of adenine-nucleotide coenzymes.

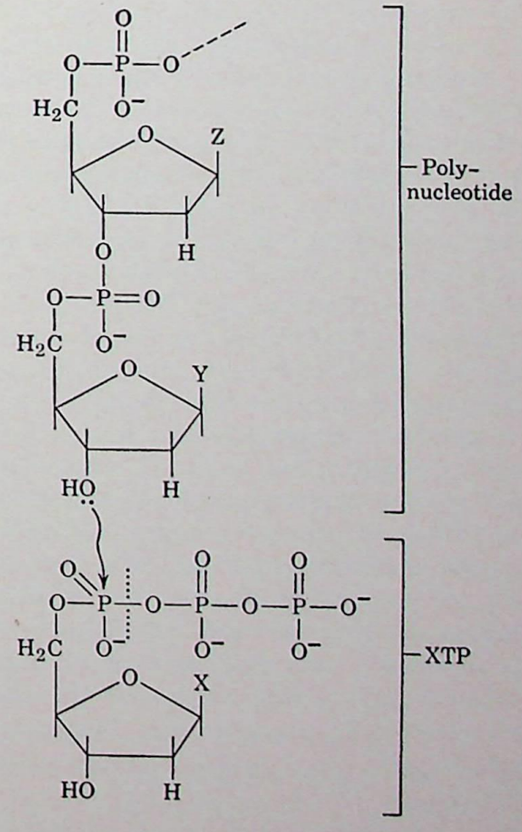


FIG. 6. Postulated mechanism for extending a DNA chain.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

TABLE I. Enzyme purification.

Fraction number	Step	Units[a] per ml	Units[a] Total	Protein mg per ml	Specific activity units per mg protein
I	Sonic extract	2.0	16 800	20.0	0.1
II	Streptomycin	13.0	19 500	3.0	4.3
III	DNase, dialysis	12.1	18 100	1.80	6.7
IV	Alumina gel	15.4	12 300	0.78	20.
V	Concentration of gel eluate	110.	9 000	4.90	22.
VI	Ammonium sulfate	670.	6 030	8.40	80.
VII	DEAE resin	120.	3 600	0.60	200.–400.

[a] A unit is defined as the amount of enzyme causing the incorporation of 10 mμmoles of thymidine triphosphate into an acid-insoluble fraction during the assay period of 30 min at 37°. See Table II for further details about the assay incubation mixture.

involved in a nucleophilic attack on a deoxynucleoside triphosphate, thus extending the chain by one unit and displacing pyrophosphate.

We approached the problem of nucleic-acid synthesis experimentally by mixing together and incubating the following: a deoxynucleoside labeled very intensely in one of the carbon atoms of its base, an extract made from an exponentially growing culture of *E. coli* (generation time 20 min), ATP as an energy source, and Mg ions. To determine whether any nucleic-acid synthesis had taken place, one relied upon the very simple and fortunate chemical property of DNA that it is precipitated at acid *p*H, while the deoxynucleotides which were the substrates are completely soluble under these conditions. It was ascertained many times that thorough washing of the DNA precipitate removed even the smallest traces of the deoxynucleoside substrates. Thus, in our earliest experiments, traces of radioactive thymidine (thymine deoxyriboside) were found incorporated into an acid-precipitable substance that could be rendered *non*precipitable by treatment with purified deoxyribonuclease (DNase).[18] In this crude system, it was possible to detect 50 counts in the precipitate out of 5 000 000 counts in 1 μmole of the starting substrate. This indicated an incorporation of only 10 μμmoles or about 1/10 000 of the amount detectable by the most sensitive microchemical methods.

Using the incorporation of C^{14}-thymidine into a DNase-sensitive acid-precipitable material as the assay, purification was begun of the components of the *E. coli* extract essential for this reaction. It soon was recognized that one of the enzymatic functions of the crude fraction was the conversion of the thymidine through thymidylate to the triphosphate level, and thymidine triphosphate was used thereafter as the substrate.[26,27]

The enzyme-purification procedure at present[28] involves first the preparation of an extract of *E. coli* by sonic treatment, then several fractionation steps (Table I) resulting in an enrichment with respect to protein of about 2000 to 4000 times over the crude extract. With this preparation, certain interesting features of the reaction have become apparent and are discussed presently.

Even at an intermediate stage of purification, it became clear that, in order for one of the triphosphates to be incorporated into DNA, the triphosphates of all four of the deoxynucleosides which commonly occur in DNA, and also DNA itself, must be present.[29] This is illustrated in Table II. The incubation mixture contained the triphosphates* of thymidine, deoxycytidine, deoxyguanosine, and deoxyadenosine, calf-thymus DNA, Mg^{++}, and enzyme (with the most purified of the enzyme preparations, about 0.1 μg of protein was used). The isotopic marker used has been either C^{14} in carbon-2 of thymine or P^{32} in the innermost phosphate group of one of the triphosphates. Under these conditions, 0.5 mμmole of the labeled nucleotide was incorporated. Omission of any one of the triphosphates, of the DNA, or of Mg^{++}, or pretreatment of the DNA with crystalline pancreatic DNase reduced the reaction to a level below the limits of detectability. Replacement of the triphosphates by the corresponding deoxynucleoside diphosphates reduced the reaction to or near the limits of detectability (<5%). Nor was there any detectable reaction when ribose analogs such as ATP or CTP were used in place of dATP or dCTP, respectively. Also, there was no reaction when DNA was replaced by RNA or by DNA degraded by acid treatment or sonic irradiation. However, as is mentioned later, DNA from a variety of sources, plant, animal, and virus, served in the reaction.

The enzymatic incorporation of deoxynucleotide into a DNA fraction using incubation conditions similar to those in Table II also has been observed with the use of sonic extracts of other bacterial species (*Hemophilus influenzae*, *Aerobacter aerogenes*), and with extracts of several types of animal cells (HeLa cell cultures, lymph-gland and leukemic cells).[29] While assays for the DNA-

TABLE II. Requirements for deoxynucleotide incorporation into DNA.

	mμmoles
Complete system[a]	0.50
Omit dCTP, dGTP, dATP	<0.01
Omit dCTP	<0.01
Omit dGTP	<0.01
Omit dATP	<0.01
Omit Mg^{++}	<0.01
Omit DNA	<0.01
DNA pretreated with DNase	<0.01

[a] The complete system contained 5 mμmoles of TP^{32} PP (1.5×10⁶ cpm per μmole), dATP, dCTP, and dGTP, 1 μmole of $MgCl_2$, 20 μmoles of glycine buffer (pH 9.2), 10 γ of calf-thymus DNA, and 3 γ of enzyme Fraction V, in a final volume of 0.30 ml. The incubation was carried out at 37° for 30 min.

DNase treatment of DNA was carried out in an incubation mixture containing 60 γ of thymus DNA, 50 μmoles of Tris buffer, *p*H 7.5, 5 μmoles of $MgCl_2$, and 5 γ of pancreatic DNase, in a final volume of 0.50 ml. After 30 min at 37°, 0.02 ml of 5% bovine serum albumin and 0.05 ml of 5*N* perchloric acid were added. The precipitate was centrifuged and dissolved in dilute alkali and neutralized. A control incubation from which DNase was omitted yielded fully active DNA.

* For the triphosphates of thymidine, deoxycytidine, deoxyguanosine, and deoxyadenosine, respectively, the following abbreviations are used: TTP, dCTP, dGTP, and dATP. The symbol $TP^{32}PP$ indicates that the triphosphate is labeled with P^{32} in the innermost phosphate group.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

synthesizing system cannot be regarded as accurate in the crude extracts, it is significant that the values for the bacterial extracts are of the order of 20 to 50 times greater than those for the animal-cell extracts. Bollum and Potter[31,32] have reported the incorporation of H^3-thymidine into DNA by extracts of regenerating rat liver.

With the most purified *E. coli* enzyme preparation, it has become possible to demonstrate net synthesis by the use of more-direct chemical methods. Several experiments are illustrated in Table III. In experiment 1, there was an increase in DNA of a little over twofold, measured by spectrophotometry, deoxypentose assay, or by incorporation of radioactive tracer. In the other experiments, increases of DNA by a factor of 10 to 20 were obtained, and 90 to 95%, therefore, of isolated DNA was derived from the deoxynucleotide substrates supplied. The factors responsible for cessation of the reaction are currently under study.

On the basis of the foregoing results, one can consider the following over-all equation for the enzymatic synthesis of DNA:

$$\left.\begin{matrix} n\ \mathrm{TPPP} \\ n\mathrm{dGPPP} \\ n\mathrm{dAPPP} \\ n\mathrm{dCPPP} \end{matrix}\right\} + \mathrm{DNA} \rightleftharpoons \mathrm{DNA} - \begin{bmatrix} \mathrm{TP} \\ \mathrm{dGP} \\ \mathrm{dAP} \\ \mathrm{dCP} \end{bmatrix}_n + 4(n)\mathrm{PP}$$

The four triphosphates+DNA yield a product which contains the four nucleotidyl residues linked in some covalent and hydrogen-bonded fashion with the DNA. It is apparent from the equation that inorganic pyrophosphate (PP) must be split out during the reaction and in quantities equimolar to the amounts of deoxynucleotide incorporated into the DNA. It might be considered also, especially in the light of the Ochoa studies[33] of the reversibility of synthesis of ribonucleic-acid polymers, that there might be a pyrophosphorolytic reversal of the reaction. It has been shown[18] that, with the incorporation of C^{14}-thymidylate into DNA, equimolar amounts of PP were released; this was isolated and identified by ion-exchange chromatography.

TABLE III. Net synthesis of DNA.

Expt. No.[a]	Estimation	Control (no enzyme)	Complete	Δ
		μmoles	μmoles	μmoles
1	P^{32} incorporation	0.00	0.28	0.28
	Optical density	0.19	0.46	0.27
	Deoxypentose	0.19	0.40	0.21
2	Optical density	0.06	0.63	0.57
3	Optical density	0.05	0.58	0.53
4	Optical density	0.05	0.64	0.59
5	Optical density	0.04	0.89	0.85

[a] In experiment 1, the incubation mixture (3.0 ml) contained 0.15 μmole of $dAP^{32}PP$ (1.3×10^6 cpm per μmole), 0.3 μmole of dGTP, 0.15 μmole of dCTP, 0.15 μmole of dTTP, 200 μmoles of potassium-phosphate buffer (*p*H 7.4), 20 μmoles of $MgCl_2$, 0.1 mg of calf-thymus DNA, and 12 γ of enzyme Fraction VII. The mixture was incubated at 37° for 180 min. DNA was precipitated, washed, taken up in 1.2 ml of 0.5*N* perchloric acid, and heated for 15 min in a boiling water bath. Optical-density measurements were made at 260 mμ and converted to nucleotide equivalents using a molar extinction coefficient of 8960 (derived from the calculated values for an acid hydrolyzate of calf-thymus DNA). In the P^{32} estimation of DNA synthesis, incorporation of deoxyadenylate was multiplied by a factor based on its percentage in calf-thymus DNA. The radioactivity actually observed for the controls did not exceed the background count. In experiments 2, 3, 4, and 5, the reaction mixture (1.0 ml) contained 0.32 μmole of each of the four triphosphates, 30 γ of calf-thymus DNA, 60 μmoles of potassium-phosphate buffer (*p*H 7.4), 6 μmoles of $MgCl_2$, and 8 γ of enzyme Fraction VII. The mixture was incubated for 250 min at 37°. 2*M* NaCl then was added to give a final concentration of 0.2*M* and the mixture was heated for 5 min at 70°. Unreacted triphosphates were removed by exhaustive dialysis against 0.2*M* NaCl. The product contained no acid-soluble material. Optical density at 260 mμ was determined and converted to nucleotide equivalents using a molar extinction for DNA of 6900.

TABLE IV. Physical properties of DNA.

	Primer	Product	Heated at 100°C, 15 min Primer	Heated at 100°C, 15 min Product
Sedimentation coefficient	25	20–25	20	14
Intrinsic viscosity	50	15–35	<1	<1
Molecular weight	8×10^6	$4\text{–}6\times10^6$	$<1\times10^6$	$<0.7\times10^6$

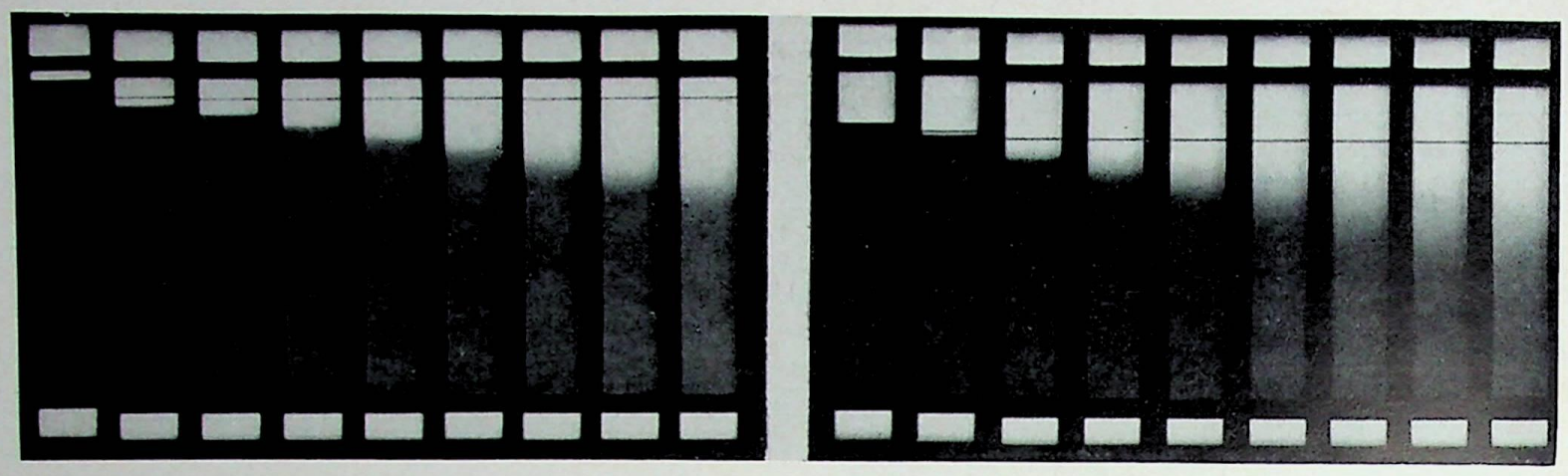
Evidence for reversal of the reaction has been obtained.[29] When PP is present in concentrations of 2×10^{-3} M—i.e., about 100 times the concentration of the deoxynucleoside triphosphates—the synthetic reaction is inhibited by about 50%. Under such conditions, when PP^{32} is used, its incorporation into the terminal PP groups of the four triphosphates has been observed. The rate of the reaction is comparable to the synthetic rate. The reaction is absolutely dependent upon the presence of DNA. DNA degraded with DNase is inert. Inorganic orthophosphate fails to replace PP. A distinctive feature of the reaction is that triphosphates are required. When they are omitted, there is only a small reaction, which may be interpreted as a very limited pyrophosphorolysis of DNA. It is necessary to have only one triphosphate to augment the pyrophosphorolysis reaction considerably.

Up to this point, the discussion has been concerned with how a purified enzyme preparation can synthesize, in the presence of four deoxynucleoside triphosphates and of DNA, presumably acting as primer, a product that is insoluble at acid *p*H and which is nondialyzable and may be degraded to acid-soluble fragments by the action of pancreatic DNase. It has been established with more-sensitive techniques that even a single triphosphate may react, but to an extent so limited as to suggest the addition of one or very few residues to the ends of the DNA chains added as acceptor.[34]

A more detailed consideration is presented below of the physical and chemical nature of the enzymatic product, DNA, prepared under conditions where 90% or more of the polymerized material is newly synthesized.

In physicochemical studies, we have been able to show that the enzymatically synthesized product has essentially the same physical characteristics as DNA carefully prepared from calf thymus.[35] The sedimenta-

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA



Calf-thymus DNA
$S=24$

Enzymatically synthesized DNA
$S=24$

FIG. 7. Sedimentation of DNA primer and the enzymatically synthesized product.

tion behavior (Fig. 7) of the calf-thymus and enzymatically synthesized samples were quite similar, although the enzymatically synthesized sample showed a greater polydispersity. The latter may be the result of the action of contaminating DNases in the enzyme preparation. The sedimentation constants in this and in a great many other runs have been in the neighborhood of 25 to 30. Viscosimetric determinations also have yielded values of 15 to 35 deciliters/gram which are comparable to those obtained for calf-thymus DNA (40 to 50); from these, average molecular weights of 4 to 6 million for several samples of enzymatically synthesized DNA were calculated (Table IV). It may be inferred from the sedimentation and viscosity characteristics of the product that it consists of highly ordered rigid structures with effective volumes greater than would be expected from single polynucleotide chains with freedom of rotation at each link in the backbone. In support of this view was the collapse of the macromolecular structure when the DNA was heated for 15 min at 100°C; like thymus DNA, the viscosity decreased to immeasurably low levels, while the sedimentation rate decreased only slightly. Furthermore, there was found a typical hyperchromic effect[36] upon degradation of the product with DNase. The curves in Fig. 8 show that just as there is an increase in optical density upon digestion of calf-thymus DNA with pancreatic DNase, so is there a kinetically similar increase with the enzymatic product and to the same extent of about 30% above the starting values.

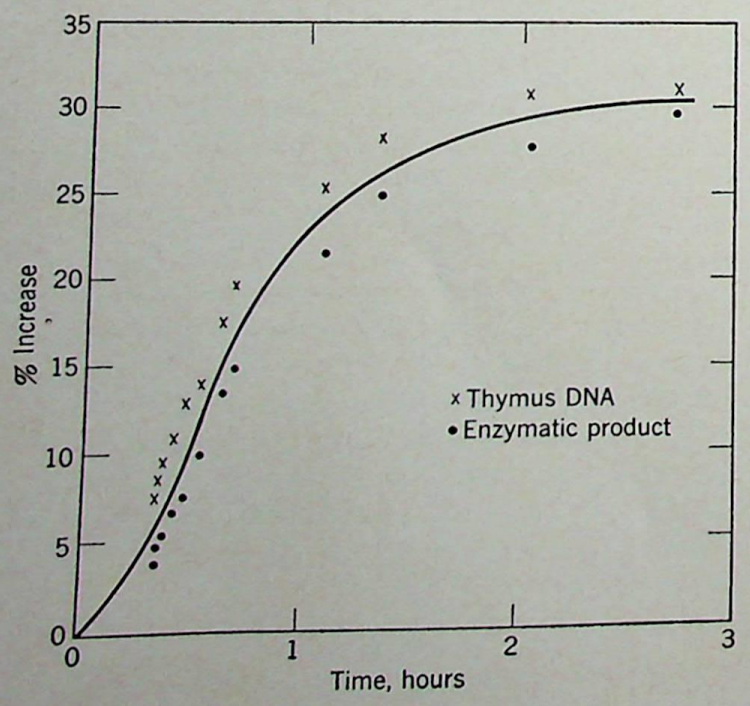


FIG. 8. Increase in ultraviolet absorption of DNA upon digestion with pancreatic DNase.

One now comes to a consideration of the chemical structure of the product. To begin with, it can be affirmed that the substrates are linked in the DNA product by typical 3′–5′ diester bonds.[29] What can be said of the base composition and the ratios of bases of the product? Do they bear any relation to the DNA primer added to the reaction?

In Table V are the base-composition data on enzymatically synthesized DNA. A close correspondence exists in the enzymatically synthesized DNA's between the contents of adenine and thymine on the one hand and of guanine and cytosine on the other, so that the ratio of purines to pyrimidines (A+G/C+T) is in each case nearly unity. Furthermore, good agreement exists for the ratio (A+T/G+C) between the enzymatic product and its corresponding primer, ranging from values of 0.59 for *Mycobacterium phlei* to greater than 40 for an enzymatically synthesized copolymer of deoxyadenylate and thymidylate. The latter is formed by the DNA-synthesizing enzyme in the absence of primer under rather specialized, but poorly understood, conditions: specifically, after lag periods of from 3 to 6 hr. Once formed, such a polymer then can be replicated without a lag and will consist solely of adenine and thymine, although all four of the deoxynucleoside triphosphates are provided in the reaction mixture. This polymer, therefore, represents an extreme case in which the base composition of the enzymatic product reflects that of the primer.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

TABLE V. Purine and pyrimidine composition of enzymatically synthesized DNA.[a,b]

DNA		Number of analyses	A	T	G	C	$\frac{A+T}{G+C}$	$\frac{A+G}{T+C}$
M. phlei	primer	3	0.65	0.66	1.35	1.34	0.49(0.48–0.49)	1.01(0.98–1.04)
	product	3	0.66	0.80	1.17	1.34	0.59(0.57–0.63)	0.85(0.78–0.88)
A. aerogenes	primer	1	0.90	0.90	1.10	1.10	0.82	1.00
	product	3	1.02	1.00	0.97	1.01	1.03(0.96–1.13)	0.99(0.95–1.01)
E. coli	primer	2	1.00	0.97	0.98	1.05	0.97(0.96–0.99)	0.98(0.97–0.99)
	product	2	1.04	1.00	0.97	0.98	1.02(0.96–1.07)	1.01(0.96–1.06)
Calf thymus	primer	2	1.14	1.05	0.90	0.85	1.25(1.24–1.26)	1.05(1.03–1.08)
	product	6	1.19	1.19	0.81	0.83	1.46(1.22–1.67)	0.99(0.82–1.04)
T2 phage	primer	2	1.31	1.32	0.67	0.70	1.92(1.86–1.97)	0.98(0.95–1.01)
	product	2	1.33	1.29	0.69	0.70	1.90(1.82–1.98)	1.01(1.01–1.03)
"Synthetic A-T Copolymer"	···	1	1.99	1.93	<0.05	<0.05	>40.	1.05

[a] A, T, G, and C refer, respectively, to adenine, thymine, guanine, and cytosine, except that C in the case of *T2*-phage primer refers to hydroxymethylcytosine.
[b] The figures in parentheses represent the range of values obtained.

The somewhat higher values of (A+T/G+C) observed for some of the products most probably can be attributed to contamination of these DNA's with traces of this deoxyadenylate-thymidylate copolymer. The obvious implications of these early results are that the added DNA is serving as a template for an enzymatic replication of DNA, but it is evident too that more-extensive documentation is necessary before this conclusion can be considered to be established.

In reviewing the specificity of this DNA-synthesizing system, the results have indicated that samples of DNA from a variety of origins can serve as primers. It has been mentioned also that only the *tri*phosphates of the *deoxy*nucleosides are reactive. What can be said of the specificity of the substrates with respect to the structure of the pyrimidine and purine bases? From the many interesting reports on the incorporation of bromouracil,[37–39] of azaguanine,[40] and of other analogs into bacterial and viral DNA, it might be surmised that some latitude in the structures of the bases can be tolerated provided there is no interference with their hydrogen bonding. It would be well to reiterate at once what Rich mentioned earlier (p. 191). Analysis of the composition of samples of DNA from a great variety of sources and by many investigators (reviewed by Chargaff[41]) reveals the remarkable fact that the purine content always equals the pyrimidine content. Among the purines, the adenine content may differ considerably from the guanine, and among the pyrimidines, the thymine from the cytosine. There is an invariable equivalence, however, between bases with an amino group in the 6 position of the ring, such as in adenine and cytosine, and the bases with a keto group in the 6 position of the ring, such as in guanine and thymine. These facts[41,42] were interpreted on the basis of hydrogen bonding by Watson and Crick[43] in their masterful hypothesis on the structure of DNA. In a given species, the DNA composition of all of the cells is characterized by a distinctive ratio of the number of adenine-thymine pairs to the number of guanine-cytosine pairs.

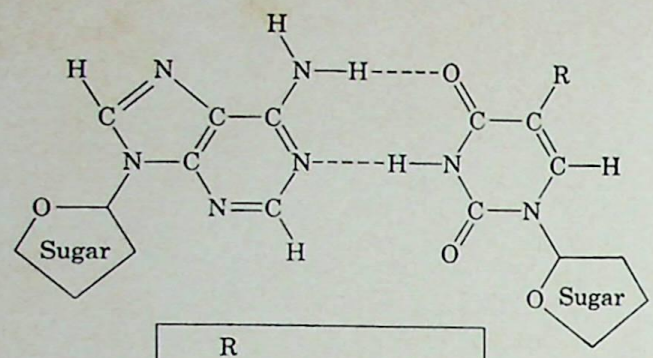
TABLE VI. Replacement of natural bases by analogs in enzymatic synthesis of DNA.

Expt. No.	Control value[a] (mμmoles)	Base analog used	Natural base omitted (Percent of control)[b] Thymine	Adenine	Guanine	Cytosine
1	0.50	Uracil	*54*	4	6	
1a	0.88	Uracil				3
2	0.43	5-bromouracil	*97*	2	4	
2a	0.42	5-bromouracil				4
3	0.51	5-bromocytosine		4	4	*118*
3a	0.40	5-bromocytosine	4			
4	0.58	5-methylcytosine		2	3	*185*
4a	0.52	5-methylcytosine	2			
5	0.37	Hypoxanthine		3	*25*	5
5a	0.27	Hypoxanthine	4			

[a] Control values are mμmoles of radioactive deoxynucleotide incorporated into DNA in the absence of analog. Incubation mixtures contained in 0.3 ml, 5 mμmoles each of TTP, dATP, dCTP, and dGTP; 2 μmoles of $MgCl_2$; 20 μmoles of potassium phosphate (pH 7.4); 10 μg of calf-thymus DNA; and 1μg of enzyme fraction VII-R. Experiments were performed at 37° for 30 min. Labeled substrates were: $dCP^{32}PP$ in Expts. 1, 2, 5a; $TP^{32}PP$ in Expts. 1a, 3,4,5; and $dGP^{32}PP$ in Expts. 2a, 3a, 4a.
[b] The percentage value represents the fraction of the labeled substrate incorporated when the analog (5 mμmoles) was used instead of a natural base. All bases, natural or analog, were supplied as the deoxynucleoside triphosphates. Values of 5% or below are near the limit of detectability and are of questionable significance.

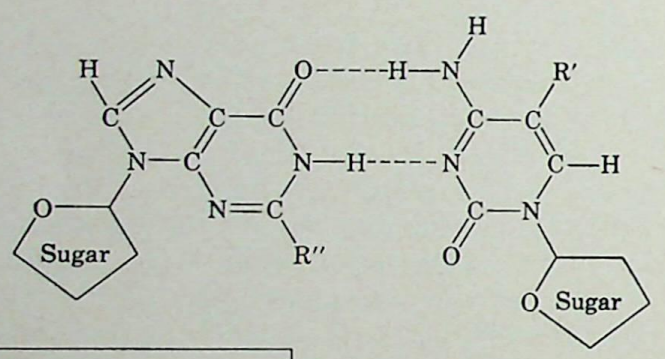
CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

Hydrogen Bonding of Adenine to Thymine



R	
—CH_3	Thymine
—H	Uracil
—Br	Bromuracil

Hydrogen Bonding of Guanine to Cytosine



R′	
—H	Cytosine
—CH_3	Methylcytosine
—Br	Bromcytosine

R″	
—NH_2	Guanine
—H	Hypoxanthine

FIG. 9. Hydrogen bonding of the bases.

As can be seen from Table VI,[44] deoxyuridine triphosphate used in place of thymidine triphosphate supported DNA synthesis at 54% of the rate of the control value but failed to support synthesis when used in place of the triphosphates of deoxyadenosine, deoxyguanosine, or deoxycytidine. 5-bromodeoxyuridine triphosphate was more effective as a replacement for thymidine triphosphate but was unable to substitute for any of the other triphosphates. The 5-bromo- and 5-methyl-deoxycytidine triphosphates replaced only deoxycytidine triphosphate and, for unexplained reasons, were even more effective than deoxycytidine triphosphate itself. Deoxyinosine triphosphate permitted a reduced rate of DNA synthesis in the absence of deoxyguanosine triphosphate, but essentially no synthesis when any one of the other triphosphates was absent.

Thus, uracil and 5-bromouracil specifically replaced thymine; 5-methyl- and 5-bromo-cytosine replaced cytosine, and hypoxanthine substituted for guanine; xanthine was not incorporated into DNA. As seen from Fig. 9, the specific replacement of the natural bases by these analogs is consistent with and offers additional support for the base-pairing relationships in the double helix proposed by Watson and Crick for the structure of DNA.

The current status of knowledge of the biochemical aspects of RNA and DNA synthesis has been sketched in the foregoing. With respect to RNA, the enzymatic information available is inadequate to explain the metabolic behavior of cells and tissues. In the case of DNA, the enzymatic studies of replication can be reconciled with genetic phenomena, but much remains to be clarified in the mechanism of the reaction. While the biological implications of these problems are exciting and pressing, the most immediate obstacles are the limited resources available to separate and characterize these macromolecules. New techniques are desperately needed now in the nucleic-acid field to cope with such problems as were solved in the protein field over the last 50 years.

BIBLIOGRAPHY

[1] P. C. Zamecnik, M. L. Stephenson, and L. I. Hecht, Proc. Natl. Acad. Sci. U. S. **44**, 73 (1958).

[2] P. Berg and E. J. Ofengand, Proc. Natl. Acad. Sci. U. S. **44**, 78 (1958).

[3] S. Ochoa, Federation Proc. **15**, 832 (1956).

[4] E. S. Canellakis, Biochim. et Biophys. Acta **23**, 217 (1957).

[5] E. S. Canellakis, Biochim. et Biophys. Acta **25**, 217 (1957).

[6] M. Edmonds and R. Abrams, Biochim. et Biophys. Acta **26**, 226 (1957).

[7] E. Herbert, V. R. Potter, and L. I. Hecht, J. Biol. Chem. **225**, 659 (1957).

[8] P. C. Zamecnik, M. L. Stephenson, J. F. Scott, and M. B. Hoagland, Federation Proc. **16**, 275 (1957).

[9] C. W. Chung and H. R. Mahler, J. Am. Chem. Soc. **80**, 3165 (1958).

[10] J. Hurwitz, Federation Proc. **17**, 247 (1958).

[11] A. B. Pardee and L. S. Prestidge, J. Bacteriol. **71**, 677 (1956).

[12] F. Gros and F. Gros, Biochim. et Biophys. Acta **22**, 200 (1956).

[13] A. I. Aronson and S. Spiegelman, Biochim. et Biophys. Acta **29**, 214 (1958).

[14] L. Astrachan, Federation Proc. **17**, 183 (1958).

[15] J. M. Buchanan, J. G. Flaks, S. C. Hartman, B. Levenberg, L. N. Lukens, and L. Warren, *Chemistry and Biology of Purines*, Ciba Foundation Symposium (J. and A. Churchill, Ltd., London, 1957), p. 233.

[16] G. R. Greenberg, Federation Proc. **12**, 651 (1953).

[17] A. Kornberg in *Phosphorus Metabolism*, W. D. McElroy and B. Glass, editors (The Johns Hopkins Press, Baltimore, Maryland, 1951), Vol. I, p. 392.

[18] A. Kornberg, Advances in Enzymol. **18**, 191 (1957).

[19] A. Kornberg in *The Chemical Basis of Heredity*, W. D. McElroy and B. Glass, editors (The Johns Hopkins Press, Baltimore, Maryland, 1957), p. 579.

[20] M. B. Hoagland and G. D. Novelli, J. Biol. Chem. **207**, 767 (1954).

[21] P. Berg, J. Biol. Chem. **222**, 991, 1015, 1025 (1956).

[22] A. Munch-Petersen, H. M. Kalckar, E. Cutolo, and E. E. B. Smith, Nature **172**, 1036 (1953).

[23] A. Munch-Petersen, Arch. Biochem. Biophys. **55**, 592 (1955).

[24] E. P. Kennedy and S. B. Weiss, J. Am. Chem. Soc. **77**, 250 (1955).

[25] D. E. Koshland in *The Mechanism of Enzyme Action*, W. D. McElroy and B. Glass, editors (The Johns Hopkins Press, Baltimore, Maryland, 1954), p. 608.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

[26] A. Kornberg, I. R. Lehman, and E. S. Simms, Federation Proc. **15**, 291 (1956).

[27] A. Kornberg, I. R. Lehman, M. J. Bessman, and E. S. Simms, Biochim. et Biophys. Acta **21**, 197 (1956).

[28] I. R. Lehman, M. J. Bessman, E. S. Simms, and A. Kornberg, J. Biol. Chem. **233**, 163 (1958).

[29] M. J. Bessman, I. R. Lehman, E. S. Simms, and A. Kornberg, J. Biol. Chem. **233**, 171 (1958).

[30] C. G. Harford and A. Kornberg, Federation Proc. **17**, 515 (1958).

[31] F. J. Bollum and V. R. Potter, J. Am. Chem. Soc. **79**, 3603 (1957).

[32] F. J. Bollum, Federation Proc. **17**, 193 (1958).

[33] M. Grunberg-Manago, P. J. Ortiz, and S. Ochoa, Biochim. et Biophys. Acta **20**, 269 (1956).

[34] J. Adler, I. R. Lehman, M. J. Bessman, E. S. Simms, and A. Kornberg, Proc. Natl. Acad. Sci. U. S. **44**, 641 (1958).

[35] H. K. Schachman, I. R. Lehman, M. J. Bessman, J. Adler, E. S. Simms, and A. Kornberg, Federation Proc. **17**, 304 (1958).

[36] M. Kunitz, J. Gen. Physiol. **33**, 349 (1950).

[37] V. F. Weygand, A. Wacker, and H. Dellweg, Z. Naturforsch. **7b**, 19 (1952).

[38] D. B. Dunn and J. D. Smith, Nature **174**, 305 (1954).

[39] S. Zamenhof and G. Griboff, Nature **174**, 306 (1954).

[40] M. R. Heinrich, V. C. Dewey, R. E. Parks, Jr., and G. W. Kidder, J. Biol. Chem. **197**, 199 (1952).

[41] E. Chargaff in *Nucleic Acids*, E. Chargaff and J. N. Davidson, editors (Academic Press, Inc., New York, 1955), Vol. I, p. 307.

[42] G. R. Wyatt, Exptl. Cell. Research Suppl. 2, 201 (1952).

[43] J. D. Watson and F. H. C. Crick, Nature **171**, 737 (1953).

[44] J. M. Bessman, I. R. Lehman, J. Adler, S. B. Zimmerman, E. S. Simms, and A. Kornberg, Proc. Natl. Acad. Sci. U. S. **44**, 633 (1958).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

REVIEWS OF MODERN PHYSICS VOLUME 31, NUMBER 1 JANUARY, 1959

25
Approaches to the Biosynthesis of Proteins

ALTON MEISTER

Department of Biochemistry, Tufts University School of Medicine, Boston 11, Massachusetts

IT is evident that, in considering protein synthesis, one must think about amino acids. Figure 1 provides a very general summary of what is known about the metabolic transformation of the amino acids. The figure shows that several hundred separate metabolic steps that do not lead to protein synthesis have been recognized. Each of these steps probably requires at least one specific protein. These enzymes, many other enzymes, and still other proteins that do not exhibit specific catalytic properties, are formed by processes about which very little is known—these are represented by a single large upward arrow in Fig. 1. All of the other arrows represent "pathways of metabolism," a series of stepwise processes. It is known that protein molecules of enormous complexity are synthesized by living cells at relatively rapid rates. Partly for this reason, there is a tendency to regard protein synthesis as a rapid spontaneous condensation of amino acids on a template. It is reasonable to believe that some type of model or template is needed for the synthesis of a specifically organized macromolecule composed of about 20 different building blocks. It is also clear that the specificity of proteins must ultimately be determined by genetic information. Nevertheless, it is highly improbable that the amino acids are brought together simultaneously by a process in which there are no intermediates. Failure to detect intermediates (peptide or other) in protein synthesis—or in intracellular protein degradation (which is often thought to be a stepwise process)—may probably be ascribed to inadequacies in the experimental approaches employed.

A complete description of the process of protein synthesis must include the mechanism of peptide-bond synthesis, and also the manner in which the amino-acid building blocks are arranged in specific sequences. Such a description should also explain the processes responsible for the formation of the specific configurations and linkages of the peptide chains of protein and for the binding to proteins of a variety of low molecular-weight compounds. In short, it is necessary to understand how essentially the same building blocks are used by living cells to yield such widely different structures as the contractile protein of muscle and the proteins of the blood, milk, silk, or hair. An attempt is made here to review some approaches to the problem.

STUDIES ON INTACT ORGANISMS, TISSUES, AND CELLS

Several general aspects of the mechanism of protein synthesis are apparent from studies on intact organisms. For example, nutritional studies on rats showed that growth was greater when all of the dietary essential amino acids were available to the animal at the same time.[1,2] If some of the essential amino acids were fed several hours after the others had been fed, no growth or less growth was observed, as compared with animals that received all of the essential amino acids at the same time. The necessity for the simultaneous availability of the amino acids for growth leads to at least two important conclusions concerning protein synthesis. First, it is evident that the amino acids are not stored in the body to an appreciable extent, but are removed by degradative reactions or by excretion. Second, it is clear that protein synthesis takes place relatively rapidly. If an adult animal is deprived of even one of the essential building blocks of proteins, there is prompt loss of appetite, decrease of dietary intake, and development of negative nitrogen balance, with eventual loss of weight. These effects are promptly reversed by administration of the missing amino acid.[3] These observations suggest that there is continual synthesis of protein in an animal that is not growing, and also that, in such an animal, there must be a balance between protein breakdown and synthesis.

Studies in which isotopic amino acids were administered to animals demonstrated that such amino acids were incorporated into the tissue proteins of growing and nongrowing animals.[4] These experiments proved the existence of an over-all dynamic equilibrium in which there was a continuous synthesis and breakdown of body protein, and made it possible to determine rates of turnover. Thus, it was observed that certain proteins (e.g., those of intestinal mucosa, liver, kidney, spleen) turn over rapidly, while in others (skin, muscle, brain) the turnover of protein amino acids is relatively slow. In a recent study,[5] Thompson and Ballou administered

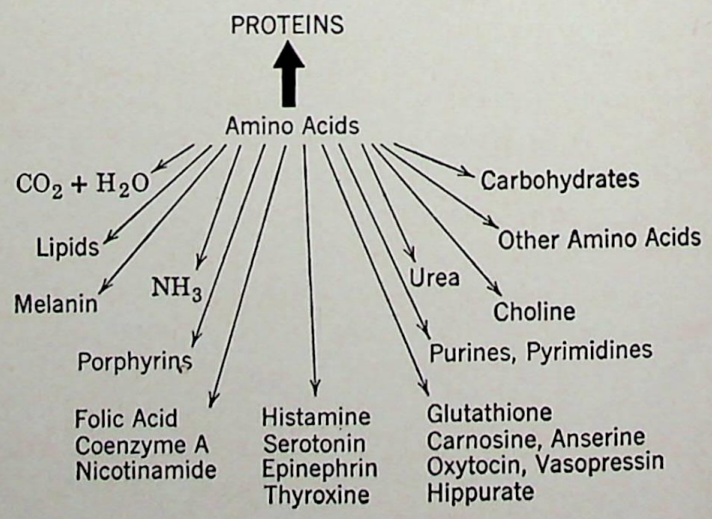


FIG. 1. General outline of amino-acid metabolism.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

tritium oxide to rats daily from conception to 6 months of age; the animals were then sacrificed at various times over a 300-day period. Also, tritium oxide was administered for 124 days to a group of mature rats, which were then sacrificed at various times over a period of 360 days. It was found that components with half-lives of several days constituted a very small proportion of the total animal, and that about half of the body components exhibited half-lives of greater than 100 days. About three-quarters of the collagen fraction exhibited an apparent half-life of 1000 days, in general agreement with other studies.[6] Collagen and the muscle proteins represent a major fraction of the total body proteins; therefore, a relatively small fraction of the body protein is extensively involved in the dynamic state. However, this fraction includes proteins that are being degraded and synthesized at remarkably rapid rates. The observed turnover phenomena may be explained in part in terms of cell destruction and cellular protein secretion. Thus, the intracellular protein molecules may be stable until they are secreted by the cell or until the cell is destroyed. The latter interpretation applies to the hemoglobin of the red blood cells, which does not turn over until cellular destruction occurs.[7]

In *E. coli* cells synthesizing adaptive enzymes, the new proteins are formed almost entirely from free amino acid in the medium, and there is practically no utilization of the amino acids of the preformed proteins.[8–10] Synthesis of induced enzymes in *P. saccharophilia*, however, was accompanied by some utilization of preformed protein in growing cells, and a larger proportion was used in resting cells.[11] In nongrowing *E. coli* cells, considerable (4 to 5% per hour) degradation and synthesis of protein has been demonstrated.[12] It appears that, in growing *E. coli* cells, the rate of protein synthesis is very rapid as compared with the rate of degradation and, therefore, that the synthesis of new protein does not involve the utilization of significant quantities of amino acids derived from the degradation of other cellular proteins. However, in *E. coli* cells that are not carrying out net synthesis of protein, considerable degradation and synthesis of protein can be demonstrated.

In studies on Ehrlich's ascites carcinoma cells, Moldave[13,14] demonstrated an uptake of labeled amino acids into the cellular proteins. Incubation of labeled cells in a medium containing C^{12}-amino acids resulted in intracellular release of radioactive amino acids without concomitant net loss of protein. Similar results were obtained by Piez and Eagle[15] in experiments on HeLa cells grown in tissue culture. It was shown that there was no appreciable reutilization of degraded cells or of secreted proteins. The available data do not indicate whether all of the individual protein molecules of the cell turn over or whether they turn over at similar rates. It is also difficult to establish or to exclude the existence or exchange, i.e., replacement, of a single amino acid or a group of amino acids of a peptide chain without complete synthesis of the protein. Such exchange might occur at an intermediate stage of protein synthesis. Exchange of an amino acid with an amino-acid residue in the interior of a protein would require opening two peptide bonds and a mechanism that would bring the two portions of the peptide chain together after acceptance of the new amino-acid molecule. That some of the steps of protein synthesis may be reversible must be considered in connection with the turnover and exchange phenomena considered above. Studies by Simpson[16] showed that the release of amino acids from rat-liver slices is depressed by conditions (anaerobiosis, 2,4-dinitrophenol, cyanide) that inhibit utilization of energy. Amino-acid incorporation is also reduced under these conditions. The processes of degradation and synthesis may be closely related and could perhaps represent separate aspects of the same mechanism.

It is evident that information concerning the chemical reactions leading to protein synthesis will be needed to interpret the observed exchange phenomena. However, experimental work bearing on this question has arisen from studies designed to determine whether protein synthesis takes place by a stepwise process (in which free peptides or other peptide intermediates are formed), or by a mechanism in which the amino acids are arranged at one time on a template corresponding to the sequence of the protein to be synthesized and then dislodged from the template in a single step. The "template" idea is attractive; it provides a mechanism for a specific amino-acid sequence and is consistent with the "all-or-none" aspects of nutritional studies. Experimental studies relating to this problem fall in two categories: those which show that labeled amino acids are incorporated to the same extent into different positions of the peptide chains of proteins, and those which show that the labeling is unequal, i.e., that the specific activities of amino acids obtained from different portions of the protein molecule are not the same. The studies of Muir *et al.*[17] fall into the first category. They found that the *N*-terminal valine and the total of the nonterminal valine residues of the hemoglobin of rats previously injected with C^{14}-valine exhibited the same specific activity. Askonas *et al.*[18] injected a lactating goat with labeled amino acids and found that the specific activities of amino acids obtained from different portions of the casein molecule were, within experimental error, the same. On the other hand, Steinberg *et al.*[19] labeled the ovalbumin of hen's oviduct by incubating this tissue with $C^{14}O_2$, and found that the specific activity of the aspartate of the hexapeptide (cleaved by enzymatic conversion of the ovalbumin to plakalbumin) was significantly greater than that of the plakalbumin. Similar results were obtained in *in vivo* experiments, and with other proteins. The finding of unequal labeling is consistent with the existence of peptide intermediates, but other interpretations are not excluded. Uniform labeling would be the expected result if amino acids were added at one time to a tem-

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

plate. Uniform labeling might result with peptide intermediates also, if these intermediates equilibrated rapidly with the available free amino acids. The recent discovery of separate amino-acid activating enzymes and ribonucleic-acid acceptors (see p. 217) makes it possible to conceive of mechanisms by which amino acids may be incorporated into different types of intermediates for protein synthesis. On the other hand, it is possible that templates exist for certain portions of protein molecules and that, in the final stages of protein synthesis, one or more large peptides prepared on different templates are combined. For example, the *A*- and *B*-peptide chains of insulin might require separate templates, and, if the two chains were not synthesized at the same rates, or if the templates were filled by amino acids from different pools, unequal labeling of the final molecule would result. Yet, in studies in which insulin was synthesized by calf-pancreas slices in the presence of labeled amino acids, unequal labeling within both of the *A*- and *B*-chains was observed; also, unequal labeling of ribonuclease was observed.

The finding of unequal labeling is clearly more significant than that of uniform labeling. Although unequal labeling might occur if a newly synthesized, uniformly labeled protein participated in subsequent "exchange" reactions, the evidence indicates that unequal labeling is observed in studies where the interval of time between introduction of free, labeled amino acid and isolation of the protein is relatively short. With longer time intervals, in the same system, labeling becomes uniform. Although the available data do not permit unequivocal conclusions, it is clear that the unequal labeling phenomenon, which has now been observed with several proteins, must be taken into account in any complete explanation of the process of protein synthesis.

As discussed above, the studies in which labeled amino acids were administered to intact animals revealed that relatively high concentrations of isotope were found in those organs known to be capable of very active protein synthesis, such as the liver and pancreas. When these organs were separated by differential centrifugation into subcellular fractions (nuclei, mitochondria, microsomes), the highest concentrations of isotope were found in the microsomal fraction. Thus, Hultin found that injected N^{15}-glycine was taken up most rapidly by the microsomal fraction of the liver of chicks.[20] Similar results were observed in experiments with mammalian tissues.[21,22] The concept has, therefore, developed that the microsomal particles of these tissues represent the most active fraction of the cell in protein synthesis. The microsome fraction (as usually obtained by differential centrifugation of liver homogenates) includes ribonucleoprotein particles which are attached to membranous material, and perhaps also to other cellular matter.[23,24] When the microsomal particles are treated with sodium deoxycholate, most of the lipid and lipoprotein portions of the microsome are removed, leaving insoluble ribonucleoprotein particles which contain about equal weights of protein and ribonucleic acid. These particles contain most of the ribonucleic acid of the microsomal fraction and about one-sixth of the total protein of this fraction.[25] When such particles were isolated from the livers of animals previously injected with labeled amino acids, they were found to have incorporated isotope more rapidly than the lipoprotein portion of the microsome and the other subcellular fractions.[26] Similar findings have been made in different laboratories, on several tissues and in various species. The findings, which are remarkably similar, lead to the conclusion that the most rapid uptake of amino acids takes place in the ribonucleoprotein particle of the microsomal fraction. However, as indicated below, mitochondria and nuclei also incorporate amino acids.

A close relationship between nucleic acids and protein synthesis has become increasingly evident.[27,28] It has been suggested that nucleic-acid synthesis must proceed concomitantly with the synthesis of protein. This conclusion is based on experiments which have shown a close correlation between the two synthetic processes (cf. Spiegelman *et al.*[29]). Nevertheless, it must be emphasized that it is not proven that the two syntheses are interdependent, and there is evidence inconsistent with this hypothesis. For example, in studies on phosphate-starved yeast, synthesis of protein was observed, whereas there was no synthesis of RNA.[30] Synthesis of amylase in pigeon-pancreas slices was not accompanied by net RNA synthesis.[31] In view of these findings, it is difficult to accept the view that protein and nucleic-acid synthesis are obligatorily linked, although there is much evidence for the belief that RNA plays a role in protein synthesis.

CELL-FREE SYSTEMS

Although important concepts concerning protein synthesis have arisen from studies on intact organisms and tissue slices, a detailed explanation of protein synthesis requires isolation and study of the cellular catalytic components. Several approaches have been made to the study of protein synthesis in cell-free systems. One of these concerns enzyme systems, initially recognized because of their ability to catalyze hydrolysis of peptide bonds. Another line of investigation has been directed toward study of certain "model" systems that synthesize peptide or pseudo-peptide bonds. Finally, attempts have been made to study incorporation of isotopically labeled amino acids into proteins in cell-free systems. This work has been facilitated by the studies on model systems and by information obtained from cytological investigations.

Formation of New Peptide Bonds Catalyzed by "Hydrolytic" Enzymes

The ability of proteolytic enzymes to catalyze the synthesis of peptide bonds was observed first more than

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

60 years ago. In these early experiments and in later ones, partial hydrolyzates of proteins were incubated with proteolytic enzymes; the formation of insoluble, high molecular-weight (2000 to 400 000) polymers (plasteins) occurred.[32,33] Although the free energy change associated with the formation of plasteins must be relatively small, the hydrolysis of a dipeptide to free amino acids proceeds spontaneously and virtually to completion. Significant reversal of such a reaction could probably not occur at the concentrations of amino acids usually present in the cell, unless a mechanism existed for continuous removal of the peptide from solution. The free energy change for the formation of glycylglycylglycylglycine from glycylglycine is about half of that for the formation of glycylglycine from glycine.[34] Thus, it appears that the free energy change for the formation of small peptides from amino acids is much greater than that for condensation of peptides to form larger peptides. If this type of reaction plays a role in the synthesis of protein, it would be expected to be significant later in the process when relatively large molecules have been formed by other reactions.

The recognition that proteolytic enzymes could catalyze replacement reactions and thus form new peptide bonds has led to a number of studies on transpeptidation reactions.[35–37] Fruton and collaborators[35,36] discovered that cathepsin-*C* (which hydrolyzes certain dipeptides at *p*H values near 5) catalyzes polymerization reactions at *p*H 7.5. Thus, glycyl-L-phenylalaninamide reacts to form an octapeptide amide as follows:

$$\text{gly-phe-NH}_2 + \text{gly-phe-NH}_2 \downarrow$$

$$\text{gly-phe-gly-phe-NH}_2 + \text{NH}_3 \quad | \text{ gly-phe-NH}_2 \downarrow$$

$$\text{gly-phe-gly-phe-gly-phe-NH}_2 + \text{NH}_3 \quad | \text{ gly-phe-NH}_2 \downarrow$$

$$\text{gly-phe-gly-phe-gly-phe-gly-phe-NH}_2 + \text{NH}_3$$

The reaction is similar in some respects to the formation of amylose from glucose-1-phosphate and also to the enzymatic synthesis of polynucleotides from nucleoside diphosphates.

γ-Glutamyl transpeptidation reactions may be involved in the synthesis of the polyglutamic acid produced by certain bacteria.[38–40] An enzyme preparation from *B. subtilis* catalyzes a reaction in which the γ-glutamyl group of glutamine is transferred to D-glutamic acid or to α-D-glutamyl-D-glutamic acid to yield di- and tri-peptides. Continuation of this process could yield polyglutamic-acid molecules of considerable size. Accordingly, the glutamine-synthesis system and the transpeptidation enzyme would be the major catalytic components required for the synthesis of polyglutamic acid from glutamic acid.

It has recently been shown that transpeptidation reactions involving proteins can take place.[41] Thus, when insulin and C^{14}-glycyl-L-tyrosinamide were incubated with cathepsin-*C*, labeled insulin could be recovered, and it was found that the labeling occurred by substitution of an *N*-terminal amino-acid residue (probably mainly the α-amino group of the *N*-terminal glycyl residue). Similarly, incubation of rat-liver mitochondria with C^{14}-tyrosinamide resulted in incorporation of isotope into the mitochondria.

Incorporation of Amino Acids into Proteins

It was shown about ten years ago that homogenates of animal tissues could incorporate labeled L-amino acids into proteins.[42] The general procedure employed involved incubation of a tissue preparation with a labeled amino acid followed by precipitation of the protein with trichloroacetic acid. Subsequent treatment of the precipitate was designed to remove nonprotein components and labeled amino acid not bound to protein. It is obvious that this approach to the study of protein synthesis has a number of inherent difficulties. For example, types of binding other than those involving peptide linkage may occur. Furthermore, the relatively small quantities of isotope incorporated and the limited amounts of labeled material available are factors that have thus far prevented isolation of pure proteins. It was subsequently recognized that incorporation was inhibited by anaerobiosis and by inhibitors of oxidation and phosphorylation,[43] and that ATP could supply the energy for amino-acid incorporation.[42–45] As in the *in vivo* experiments, incorporation studies with liver homogenates led to greater labeling of the microsomes than of the other subcellular fractions.[45] Incorporation of labeled amino acids into protein was shown to take place in a system obtained from rat-liver homogenates consisting of a microsome-rich fraction, a soluble nondialyzable fraction, and an ATP-generating system.[46] The incorporated amino acids were not released from the microsomes by subsequent incubation with unlabeled amino acids. When the soluble nondialyzable fraction required for the incorporation of amino acids into microsomes was precipitated at *p*H 5, it was less active in incorporating amino acids, but activity was restored by the addition of either guanosine diphosphate or guanosine triphosphate.[47] These studies were extended to rat-hepatoma and mouse-ascites tumor cells.[48] It is of interest that the microsomes obtained from one source were labeled by incubation in a system containing soluble fraction obtained from another. The incorporation of amino acids was additive rather than competitive; the incorporation of one amino acid was not stimulated by the addition of 17 other amino acids. It seems probable that the preparations employed contain a pool of amino acids, perhaps complete, although this does not seem to have been specifically investigated.

Although there can be little doubt that the micro-

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

somes actively incorporate amino acids, it appears that other subcellular fractions can also carry out this activity. Thus, Mirsky and collaborators[49,50] observed incorporation of amino acids into isolated calf-thymus nuclei. In certain respects, incorporation was similar to that observed with microsomal preparations. Thus, anaerobiosis and dinitrophenol inhibited incorporation. L-Alanine was incorporated, whereas D-alanine was not. Nuclei labeled by incubation with a C^{14}-amino acid did not lose significant radioactivity on subsequent incubation in a medium containing the corresponding unlabeled amino acid. Incubation with a complement of L-amino acids did not stimulate C^{14}-alanine incorporation. Treatment of the nuclei with deoxyribonuclease led to reduced incorporation; the reduction in incorporation did, in fact, parallel the removal of DNA. The decrease in incorporation observed after removal of the DNA appears to be associated with a decrease in the ability of the nuclei to synthesize ATP; thus, recent work has shown that addition of polynucleotides restores ATP synthesis.[51] Separation of the nuclei into different fractions was carried out after incorporation, and the greatest specific activity was found in a nonhistone protein closely associated with the DNA; the incorporation into histones was relatively low. A ribonucleoprotein complex that was easily extractable in *p*H 7.1 buffer was also highly labeled.

Incorporation of amino acids into the ribonucleoprotein particles of plants has been observed.[52] The features of the particulate amino-acid incorporating system of plants are similar to those of the animal-tissue systems.

Gale and collaborators have carried out an extensive study of amino-acid incorporation in disrupted bacterial cells.[53–55] In the presence of the necessary amino acids, development of certain enzyme activities in disrupted cell preparations of *Staphylococcus aureus* has been reported. Gale has concluded that when disrupted preparations of *Staphylococcus aureus* are incubated with a single amino acid and a source of energy (ATP and hexose diphosphate), incorporation occurs by an exchange between the added amino acid and protein-bound amino acids. In these experiments, incubation of the labeled preparation with the corresponding unlabeled amino acid and a source of energy led to release of the radioactive amino acid from the disrupted cell preparation. Removal of nucleic acid from the disrupted cell preparations reduced the extent of incorporation. Reactivation of the system was accomplished by the addition of DNA or RNA. Digestion of staphylococcal RNA by ribonuclease was reported to increase the ability of the RNA preparation to stimulate amino-acid incorporation. A number of active components were isolated from ribonuclease digests of staphylococcal or yeast RNA. The nature of these factors is not yet entirely clear, but it appears that certain of them are of low molecular weight and can be separated from polynucleotide material.

Simpson *et al.*[56,57] found that the rate of amino-acid incorporation into the mitochondria of muscle was of the same order of magnitude as that into the microsomes, and that isolated muscle and liver mitochondria were able to incorporate amino acids into protein.[58] Incorporation into liver mitochondria required ATP, and was increased by addition of magnesium ions and a soluble fraction obtained from rat liver. The mitochondrial-incorporation system appears similar to those for incorporation of amino acids into microsomes and thymus nuclei. One major and somewhat surprising difference, however, is that the incorporation of amino acids is increased by treatment of the mitochondria with ribonuclease. Inhibition by the nucleases is a characteristic feature of the incorporation systems of nuclei and microsomes.

Ochoa and Beljanski[59] have reported that a particulate fraction of sonically disrupted *Alcaligenes faecalis* incorporated C^{14}-L-amino acids into protein. Incorporation was dependent upon oxidative phosphorylation for the generation of ATP; however, the system was reported not to contain amino-acid activating enzymes (see below) as measured by the absence of amino-acid stimulated, pyrophosphate incorporation into ATP. The incorporation of one C^{14}-amino acid was stimulated by the addition of 18 other amino acids. They also observed small net increases of protein-nitrogen. Incorporation was decreased by treatment of the particles with *M* sodium chloride; reactivation could be obtained by addition of RNA or of a soluble enzyme ("amino-acid incorporation enzyme") that was purified from the supernatant. Incorporation was not reversible, nor was it inhibited by chloramphenicol. It is of considerable interest that incorporation did not appear to require activating enzymes of the type found in other systems.

The studies on disrupted bacterial systems have led to results somewhat different from those observed with the animal microsomal, mitochondrial, and nuclear systems. The major differences include the reversibility of incorporation of amino acids, and the observation of net protein synthesis. These phenomena are, of course, characteristic of intact cell systems. That some of the incorporation observed in studies with bacteria may represent synthesis of cell-wall material is suggested by recent work. Mandelstam and Rogers[60] incubated *Staphylococcus aureus* cells with radioactive glycine, glutamic acid, or lysine in the presence of chloramphenicol. The total precipitate obtained after treatment with trichloroacetic acid or organic solvents, and (in separate experiments) the cell-wall material and protein were isolated. The data obtained indicate that the amino acids were incorporated into the cell-wall material, and that the "protein synthesis" observed in the total precipitate was, in fact, an increase in the mass of the cell wall. Incorporation into cell-wall material is not sig-

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

nificantly affected by chloramphenicol; some incorporation into preparations of disrupted staphylococci can also occur in the presence of chloramphenicol. It is obviously important to determine how much of the observed incorporation by bacterial preparations represents cell-wall or cell-membrane synthesis.

Possible Intermediates in Protein Synthesis

Clues to the mode of formation of the peptide bonds of proteins might be expected to come from studies of the synthesis of smaller and chemically characterized compounds possessing peptide or pseudo-peptide linkages, e.g., benzoylglycine (hippuric acid), glutathione, glutamine, pantothenic acid. The major outlines of the enzymatic reactions leading to the synthesis of these compounds are now known. Study of these and of similar systems has contributed significantly to the understanding of amino-acid activation and the intermediates in protein synthesis. It may, therefore, be valuable to review briefly several activation phenomena leading to formation of peptide or pseudopeptide bonds.

Studies on the enzymatic synthesis of glutathione demonstrated that the synthesis from the three amino-acid components takes place by a stepwise process:[61,62]

$$\text{L-glutamic acid} + \text{L-cysteine} + \text{ATP} \xrightleftharpoons{Mg^{++}} \text{L-}\gamma\text{-glutamylcysteine} + \text{ADP} + \text{Pi}, \quad (1)$$

$$\text{L-}\gamma\text{-glutamylcysteine} + \text{glycine} + \text{ATP} \xrightleftharpoons{Mg^{++}} \text{L-glutathione} + \text{ADP} + \text{Pi}. \quad (2)$$

Both reactions have been separately demonstrated. As yet, there is no evidence for the formation of a free intermediate in either reaction.

The first reaction in glutathione synthesis is similar to that catalyzed by the glutamine synthesis enzyme:[63–70]

$$\text{glutamic acid} + NH_3 + \text{ATP} \xrightleftharpoons{Mg^{++}} \text{glutamine} + \text{ADP} + \text{Pi}. \quad (3)$$

The purified glutamine synthesis enzyme also catalyzes synthesis of γ-glutamyl hydroxamic acid when hydroxylamine is substituted for ammonia, as well as a transfer reaction which requires Mg^{++} (or Mn^{++}) and catalytic quantities of ADP and inorganic phosphate:

$$\text{glutamine} + NH_2OH \rightleftharpoons \gamma\text{-glutamylhydroxamic acid} + NH_3. \quad (4)$$

No free intermediates have been isolated. Incubation of O^{18}-labeled glutamic acid in this system leads to the formation of O^{18}-inorganic phosphate and glutamine in stoichiometric quantities.[71,72] It is possible that the transfer of O^{18} occurs via γ-glutamylphosphate, or perhaps by a more complex mechanism. Experiments with synthetic γ-glutamylphosphate revealed that, although this compound reacted with very low concentrations of ammonia to form glutamine, there was no significant acceleration of the rate of glutamine synthesis in the presence of enzyme.[70] The possibility remains that enzyme-bound γ-glutamylphosphate is the active intermediate, according to the following scheme:

$$\text{enzyme} + \text{glutamate} + \text{ATP} \rightleftharpoons \text{enzyme-}\gamma\text{-glutamyl phosphate} + \text{ADP}, \quad (5)$$

$$\text{enzyme-}\gamma\text{-glutamyl phosphate} + NH_3 \rightleftharpoons \text{enzyme} + \text{glutamine} + \text{phosphate}. \quad (6)$$

Although the glutamine-synthesis enzyme catalyzes a transfer reaction [reaction (4)], there is thus far no evidence for the existence of a natural acceptor. Several other enzymes catalyze transfer reactions involving the γ-carboxyl group of glutamic acid, e.g., γ-glutamyl transpeptidases that act on glutathione,[37] various glutamine transferases.[73,74] With the possible exception of the γ-glutamyl transpeptidase of *B. subtilis*,[38–40] there is as yet no clue as to the physiological roles of these enzymes that act on the ω-amide or ω-carboxyl groups of the dicarboxylic amino acids. There is no evidence for γ-glutamyl or β-aspartyl linkages in proteins, although this type of linkage occurs in certain peptide antibiotics.[75] It is conceivable that activation of the dicarboxylic acids for protein synthesis occurs by primary attack on the ω-carboxyl group followed by rearrangement to an α-carboxyl-activated molecule. This type of rearrangement has been observed nonenzymatically with derivatives of glutamic acid and aspartic acid.[70] Another possible mechanism for dicarboxylic amino-acid activation is cyclic anhydride formation. Such derivatives might react enzymatically to give α-carboxyl substituted products; similar reactions occur nonenzymatically.[76]

In contrast to the activation of glutamic acid and of γ-glutamylcysteine, which are associated with a cleavage of ATP to ADP, the synthesis of pantothenic acid[77] from pantoic acid and β-alanine and of benzoylglycine from benzoic acid and glycine involves formation of pyrophosphate from ATP. The enzyme system responsible for pantothenic-acid synthesis catalyzes an incorporation of PP into ATP in the presence of pantoic acid. A similar exchange is catalyzed by the acetate-activating enzyme:

$$\text{Acetate} + \text{ATP} + \text{CoA} \rightleftharpoons \text{acetyl CoA} + \text{AMP} + \text{PP}. \quad (7)$$

Evidence for the existence of an acyl-adenylate intermediate was first achieved by Berg[78,79] in studies on the acetate-activating system. Thus, it was shown that addition of synthetic acetyl adenylate to the enzyme system could yield either ATP or acetyl CoA:

$$\text{Acetate} + \text{ATP} \rightleftharpoons \text{acetyl AMP} + \text{PP}, \quad (8)$$

$$\text{Acetyl AMP} + \text{CoASH} \rightleftharpoons \text{acetyl CoA} + \text{AMP}. \quad (9)$$

Evidence suggesting participation of acyl adenylates in activation of fatty acids was subsequently obtained.[80] Earlier, it had been shown by Chantrenne[81] that CoA

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

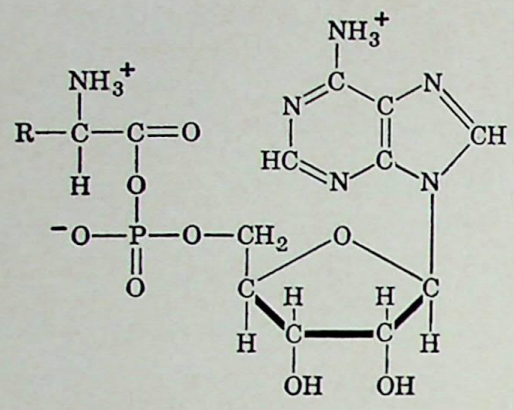


FIG. 2. General structure of aminoacyl adenylate.

was required for benzoylglycine formation. The synthesis of benzoylglycine appears to take place as follows[82–86]

$$\text{Benzoic acid} + \text{ATP} \rightleftharpoons \text{benzoyl AMP} + \text{PP}, \quad (10)$$

$$\text{Benzoyl AMP} + \text{CoA-SH} \rightleftharpoons \text{benzoyl CoA} + \text{AMP}, \quad (11)$$

$$\text{Benzoyl CoA} + \text{glycine} \rightleftharpoons \text{benzoylglycine} + \text{CoA-SH}. \quad (12)$$

Although formation of benzoyl AMP was not demonstrated, when added to the system it was active (in the presence but not in the absence of CoA-SH) in forming benzoylglycine in human liver and kidney preparations. The synthesis of phenylacetyl-L-glutamine by human tissues appears to occur by an analogous reaction.[85,86]

It thus appears that acyl adenylates are intermediates in the enzymatic activation of acetate and fatty acids, and in the synthesis of benzoylglycine and phenylacetylglutamine. The evidence for these intermediates is incomplete, however, for the enzymatic formation of these acyl adenylates has not been demonstrated. It is probable that the actual intermediates are enzyme-bound and, therefore, that the number of molecules of the intermediate present at a given time is no greater than the number of acceptor sites on the enzyme.

Amino-acid dependent, enzymatic PP-ATP exchange was independently observed by Hoagland *et al.*[87,88] and by Berg,[78,79,89] and several features of the reaction suggested that the activated amino-acid intermediate was an aminoacyl adenylate (Fig. 2). The phenomenon has now been observed and studied in a number of laboratories[90–93]; several specific amino-acid activating enzymes have been purified, and evidence has been obtained for the formation of an enzyme-bound, aminoacyl adenylate intermediate. The general reaction may be written as follows:

$$\text{Amino acid} + \text{enzyme} + \text{ATP} \xrightleftharpoons{Mg^{++}} \text{enzyme-aminoacyl adenylate} + \text{PP}. \quad (13)$$

In the presence of the specific amino acid, the enzyme catalyzes incorporation of PP into ATP. When the reaction is carried out in the presence of enzyme, ATP, Mg^{++}, amino acid, and high concentrations of hydroxylamine, the corresponding aminoacyl hydroxamate is formed. That the intermediate formed is an aminoacyl adenylate is suggested, in analogy with acetate activation, by the formation of pyrophosphate rather than orthophosphate. DeMoss *et al.*[94] found that leucyl-adenylate reacted enzymatically with pyrophosphate to give ATP (reversal of reaction[13]). Further evidence for formation of an anhydride linkage between the phosphoric-acid group of adenylic acid and the carboxyl group of amino acids was obtained in studies in which transfer of O^{18} from the carboxyl group of an amino acid to adenylic acid was observed during enzymatic activation.[95] More-direct evidence for the aminoacyl-adenylate intermediate was obtained by Karasek *et al.*,[96] who extracted a compound that exhibited the properties of synthetic tryptophanyl adenylate from reaction mixtures containing large amounts of pancreatic tryptophan-activating enzyme.[97] Also, it was found that the tryptophan-activating enzyme could catalyze ATP synthesis from pyrophosphate and a wide variety of α-aminoacyl adenylates including those of D-amino acids.[96,98] (β-Alanyl adenylate, acetyl adenylate, and carbobenzoxy α-amino acyl adenylates were not active.) On the other hand, only L-tryptophan was active in the enzymatic hydroxamate-forming reaction and in the pyrophosphate-ATP exchange. Similar results were observed with a yeast methionine-activating enzyme.[99] The explanation for this phenomenon is not yet clear. The findings with the D- and L-amino-acid derivatives are reminiscent of the observation that D-glutamic acid is activated almost as rapidly as L-glutamic acid by the glutamine-synthesis enzyme, whereas only the L-isomer of glutamine is significantly active in the transfer reaction.[67]

It would appear that the aminoacyl adenylate is stabilized by its binding to the enzyme. Such stabilization may be accomplished by a linkage involving the α-amino group; it has been found that *N*-substituted aminoacyl adenylates (e.g., carbobenzoxy-aminoacyl adenylates) are much more stable than are the corresponding free amino compounds.[100] The high reactivity of the aminoacyl adenylates is indicated by their rapid nonenzymatic reaction with hydroxylamine to form aminoacyl hydroxamates, with ammonia to yield amino-acid amides, and with amino acids to form peptides. Synthetic aminoacyl adenylates can react nonenzymatically with nucleic acid[100] and with microsomal[100] and mitochondrial[41] proteins. The incorporation of amino acids from synthetic aminoacyl adenylates into the particulate proteins and into soluble proteins represents mainly acylation of the available free amino groups of the protein. Heated mitochondrial and microsomal preparations are acylated to a much greater extent than are the corresponding unheated materials. Presumably, heat denaturation increases the number of available free amino groups. It is evident from the observations which demonstrate the high reactivity of the aminoacyl adenylates that they must be selectively transferred, and

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

..that their participation in other reactions must involve a mechanism which prevents the occurrence of nonspecific acylation reactions.

It appears that very little (if any) *free* aminoacyl adenylate is formed enzymatically. However, compounds other than hydroxylamine can react with enzymatically synthesized aminoacyl adenylates. Thus, incubation of tryptophan-C^{14} with ATP, Mg^{++}, activating enzyme, and various protein preparations (rat-liver microsomes, serum albumin, ovalbumin) gave labeled protein preparations, and similar studies with soluble RNA preparations yielded RNA that was labeled with amino acids. The reaction of enzyme-bound aminoacyl adenylates with RNA appears to be more specific, since there is evidence for the existence on soluble RNA molecules of specific amino-acid acceptor sites (see following). Aminoacyl adenylates, whether produced enzymatically or by chemical synthesis, can react with nucleotides to yield products which appear to be bound by ester linkage to the 2′ (or 3′) hydroxyl groups of the ribose moieties.

A number of specific amino-acid activating enzymes have been found. Thus, enzymes specific for methionine,[89,99] tryptophan,[97] tyrosine,[92,101] valine,[102] and leucine[101] have been described. Novelli[98] has studied the activation of amino acids by preparations of microorganisms, plants, and animal tissues. With most of these preparations, only 8 to 10 amino acids stimulated pyrophosphate-ATP exchange, although it is possible that the background exchange (owing to endogenous amino acids) may have obscured the effects of the added amino acids. With guinea-pig tissue extracts, all of the amino acids appeared to catalyze at least a small amount of exchange.[93] It is obviously of importance to determine whether or not this type of activating mechanism exists for all of the amino acids. It is curious that the tryptophan-activating enzyme should be so prominent in beef pancreas and in other sources, since this amino acid is present in very low concentrations in proteins and is absent from some proteins. However, it is possible that improved methods of isolation will be needed to obtain all of the individual enzymes.

There are difficulties associated with the study of the α-activation of glutamic acid and aspartic acid; thus, these amino acids (and their ω-amides) may form ω-carboxyl-linked intermediates. ATP exchange reactions and the hydroxamic-acid test system may, therefore, be unreliable guides as to the occurrence of α-activation. It is conceivable that these amino acids are activated by a different mechanism as suggested above.

Evidence for the accumulation of labeled intermediates during the incorporation of C^{14}-leucine into rat-liver microsomes was reported by Hultin and Beskow.[103] Hoagland, Stephenson, and Zamecnik[104,105] provided evidence for the transfer of labeled amino acids to ribonucleic acid of the soluble enzyme fraction of liver homogenate (which also contains amino-acid-activating enzyme activity). They also obtained evidence for the further transfer of the amino-acid moiety to microsomal protein. The transfer to protein was dependent upon the presence of guanosine triphosphate. Other studies have provided evidence for the hypothesis that soluble ribonucleic acid serves as an acceptor for activated amino acids.[101,102,106–108] The reaction, which appears to be analogous to other activation reactions [Eqs. (7) to (12)], but with ribonucleic acid instead of CoA as the acceptor, may be formulated as follows:

$$\text{Amino acid} + \text{ATP} \rightleftharpoons \text{aminoacyl adenylate} + \text{PP} \quad (14)$$

$$\text{Aminoacyl adenylate} + \text{RNA} \rightleftharpoons \text{amino acid-RNA} + \text{AMP}. \quad (15)$$

It appears possible that both reactions are catalyzed by the same enzyme. When the total fraction of soluble RNA is incubated with a specific amino-acid activating enzyme and the corresponding amino acid, the RNA becomes labeled. A separate activating enzyme appears necessary for linking each amino acid to RNA. There may be specific RNA sites for each amino acid; thus, labeling by several amino acids is approximately the sum of maximum labeling with each. There are also preliminary indications[101,102] that there may be separate RNA molecules for each amino acid. Formation of the RNA amino-acid complex is reversible according to Eqs. (15) and (16); thus, incubation of the RNA amino-acid complex with AMP and pyrophosphate yields ATP.[101,102,109] The RNA amino-acid complex is alkali-labile, but considerably more stable than are aminoacyl adenylates. This suggests that the linkage of amino acid to RNA does not involve an anhydride linkage. Potter and Dounce[110] isolated alkali-stable polynucleotide fractions from calf pancreas that contained amino acids and peptides; they suggested that the amino acids were bound to the nucleotides by phosphoamide bonds. The stability of this linkage to alkali effectively excludes it as the binding mechanism in the enzymatically formed RNA/amino-acid complexes. An attractive possibility, consistent with the alkali-lability of the bond, is an ester linkage involving the 2′ or 3′ hydroxyl group of ribose. Recent work by Hecht, Stephenson, and Zamecnik,[111] has shown that the soluble RNA of ascites-cell tumors incorporates adenine nucleotide into the terminal position in the presence of a soluble enzyme and ATP. Such incorporation is increased by adding cytidine triphosphate. Their evidence was interpreted to mean that cytosine and adenine nucleotides are added to RNA in this sequence. It is of interest that the incorporation of a great many amino acids was enhanced by addition of these nucleotides, suggesting that these end groups participate in the attachment of activated amino acids to RNA.

The further transfer of amino-acid moieties from RNA to microsomal protein has been described by Hoagland *et al.*, as mentioned above. The work of Hultin and Beckow,[103] which demonstrated a two-step incorpo-

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

ration of leucine into rat-liver microsomal protein, now may be interpretated in terms of the formation of an intermediate RNA-leucine complex followed by transfer to protein. The mode of transfer and the relationship of the incorporated amino acids to specific characterized proteins are major problems that await solution.

LATER STAGES OF PROTEIN SYNTHESIS

The incorporation of amino acids into the microsomal particles of animal cells has thus far not been shown to be associated with the formation of recognizable specific proteins. It is possible that the synthesis of protein molecules is completed in the ribonucleoprotein particle. Failure to detect it at this stage may be ascribed to the very small quantity of protein present, or to binding of the newly formed protein to the particle. If the latter interpretation is correct, existence of additional reactions that remove the protein and transport it to other parts of the cell must be postulated. Another explanation is that the newly formed peptide material incorporated by the microsomal particle undergoes additional transformation before becoming a recognizable protein molecule. The existence of such a protein precursor has been considered by several authors.[112,113] Such transformations might involve reactions that link relatively large peptide chains together, introduce disulfide bonds, and add smaller nonpeptide molecules. Transpeptidation reactions may function at this stage of protein synthesis. Experimental findings that appear to bear on this question have arisen from studies on the synthesis of pancreatic enzymes by Daly and Mirsky.[113] This work showed that there was no significant change in the total protein content of pancreas during secretion and synthesis; however, there were significant and marked changes in enzyme concentration. There appears, therefore, to be a distinction between rapidly formed precursor protein and specific enzyme protein. The transformation of precursor protein to specific proteins associated with secretion may be a slower process than the formation of the precursor protein. Similar conclusions are suggested by experiments of Campbell and Work[114] on incorporation of amino acids by the rabbit mammary gland, by Green and Anker[115] on the synthesis of α-globulin, and by Putnam *et al.*[116] on Bence-Jones protein synthesis. Peters[117,118] has reported that labeled carbon appears more slowly in serum albumin than in the total protein fraction, when chicken-liver slices are incubated with radioactive carbon dioxide or glycine. The appearance of isotope in the total liver-protein fraction preceded the labeling of albumin by 15 or 20 min. Further studies indicate that serum albumin, isolated by an immunological procedure, accumulates in the cytoplasmic granules before its appearance in the soluble fraction. These findings suggest that an antigenically reactive serum-albumin molecule is formed on the ribonucleoprotein particle or the associated lipoprotein membrane.

The ribonucleoprotein particles obtained from various sources have been found to contain about equal weights of RNA and protein. Such particles have been obtained from microorganisms, higher plants and animal tissues. Recent work indicates that particles obtained from the same source are not homogeneous, structurally or biochemically. Furthermore, Simkin and Work[119] found that amino acids are incorporated *in vitro* into the proteins of different microsomal subfractions at different rates, and that the pattern of *in vitro* incorporation is not the same as that observed *in vivo*.

The amino acids incorporated into isolated thymus nuclei become linked to a specific protein closely associated with the DNA.[49,50] In the mitochondrial system, incorporation of labeled valine into cytochrome-*c* was observed.[58] These findings are consistent with the belief that specific proteins may be synthesized in nuclei and mitochondria.

DISCUSSION

The available experimental data indicate that virtually all cells can incorporate amino acids, and that amino-acid incorporation (and perhaps also protein synthesis) can take place in the microsomal, nuclear, and mitochondrial fractions of certain animal cells. The general features of amino-acid incorporation into these structures are remarkably similar. The need for amino-acid activation is apparent in these systems and also in those obtained from plants and microorganisms. Additional data are required before it can be concluded that all of the amino acids are activated by formation of aminoacyl adenylates; for example, the evidence concerning the dicarboxylic amino acids and the corresponding ω-amides is incomplete, as observed above. Although acyl adenylates appear to be the reactive intermediates in acetate, fatty acid, benzoate, and phenylacetate activation, another type of activation, perhaps involving carboxyl phosphate anhydrides, appears to be involved in both steps of glutathione synthesis and in glutamine synthesis. The studies of Ochoa and Beljanski[59] lend support to the possibility that other types of amino-acid activation may occur. That phosphorylated RNA may serve as a source of energy for protein synthesis has been considered and a phosphorylated template has also been suggested. The conversion of amino acids to aminoacyl adenylates and to RNA/amino-acid complexes may form part of the mechanism responsible for transport of amino acids into the cell. Amino-acid activation may play a role in both transport and protein synthesis.

There have been many published speculations concerning the existence and nature of templates for protein synthesis. The concept of a template or model for protein formation seems almost indispensable in order to explain the high degree of specificity apparently associated with the arrangement of amino acids in proteins. A system of separate and specific activating enzymes

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

and RNA acceptor sites could provide some specificity in preparing amino acids for synthesis. Yet it appears that neither the activating systems nor the complete synthesizing system are absolutely specific, for such unnatural amino acids as ethionine and tryptazan can be activated and incorporated into protein. The incorporation of a number of "foreign" amino acids into "false" protein has now been observed. If ethionine can occasionally replace methionine in the synthesis of protein, and if fluorophenylalanine can sometimes be incorporated in place of phenylalanine or tyrosine, is it possible that the naturally occurring amino acids may themselves occasionally be "misplaced" on the peptide chains? Thus, valine might "accidentally" replace isoleucine or vice versa, and so forth. It seems possible that the specificity of amino-acid sequences may not be absolute.

The participation of soluble and particulate RNA in the process of amino-acid incorporation is consistent with the belief that RNA functions as a template. However, there is no evidence that unequivocally excludes the possibility that protein may serve as a template. Specific enzymes are apparently needed throughout the process of protein synthesis, and it appears very probable that both protein and RNA function in the formation of specific amino-acid sequences. The nature of the process by which genetic information is supplied to the protein-synthesizing mechanisms, and discussions of possible relationships between the base sequences of nucleic acids and the amino-acid sequences of proteins are considered elsewhere in this volume (see p. 227). Information relating to these important problems may arise from studies on the structure of the soluble RNA molecules that accept the amino-acid moieties of aminoacyl adenylates.

The observation that incorporation of amino acids into microsomes from one source could be carried out with soluble enzymes obtained from another is very interesting. If incorporation represents synthesis of a specific protein, then it would appear that the specificity must reside in the microsomal particle; alternatively, a "mixed" system of this type synthesizes a "hybrid" protein. It is possible that the microsomal-incorporation system represents only a portion of the cellular-protein synthetic system, and that additional components present in the intact cell must be added in order to obtain net protein synthesis. As yet, there is no unequivocal demonstration of net synthesis of a specific protein in a cell-free system. Such an achievement would be an important step toward the goal of understanding the process of protein synthesis.

BIBLIOGRAPHY

[1] L. E. Frazier, R. W. Wissler, C. H. Steffee, R. L. Woolridge, and P. R. Cannon, J. Nutrition **33**, 65 (1947).

[2] E. Geiger, Science **111**, 594 (1950).

[3] W. C. Rose, Federation Proc. **8**, 546 (1949).

[4] R. Schoenheimer, *The Dynamic State of Body Constituents* (Harvard University Press, Cambridge, 1942).

[5] R. C. Thompson and J. E. Ballou, J. Biol. Chem. **223**, 795 (1956).

[6] A. Neuberger and H. G. B. Slack, Biochem. J. **53**, 47 (1953).

[7] D. Shemin and D. Rittenberg, J. Biol. Chem. **166**, 621, 627 (1946).

[8] B. Rotman and S. Spiegelman, J. Bact. **68**, 419 (1954).

[9] D. S. Hogness, M. Cohn, and J. Monod, Biochim. et Biophys. Acta **16**, 99 (1955).

[10] A. L. Koch and H. R. Levy, J. Biol. Chem. **217**, 947 (1955).

[11] A. Markovitz and H. P. Klein, J. Bact. **70**, 641, 649 (1955).

[12] J. Mandelstam, Biochem. J. **69**, 110 (1958).

[13] K. Moldave, J. Biol. Chem. **221**, 543 (1956).

[14] K. Moldave, J. Biol. Chem. **225**, 709 (1957).

[15] K. Piez and H. Eagle, Federation Proc. **17**, 289 (1958).

[16] M. V. Simpson, J. Biol. Chem. **201**, 143 (1953).

[17] H. M. Muir, A. Neuberger, and J. C. Perrone, Biochem. J. **52**, 87 (1952).

[18] B. A. Askonas, P. N. Campbell, and T. S. Work, Biochem. J. **58**, 326 (1954).

[19] D. Steinberg, M. Vaughan, and C. B. Anfinsen, Science **124**, 389 (1956).

[20] T. Hultin, Exptl. Cell Research **1**, 376 (1950).

[21] H. Borsook, C. L. Deasy, A. J. Haagen-Smit, G. Keighley, and P. H. Lowy, J. Biol. Chem. **187**, 839 (1950).

[22] V. G. Allfrey, M. M. Daly, and A. E. Mirsky, J. Gen. Physiol. **37**, 157 (1953).

[23] G. E. Palade, J. Biophys. Biochem. Cytol. **1**, 59 (1955).

[24] F. S. Sjöstrand and V. Hanzon, Exptl. Cell Research **7**, 393 (1954).

[25] R. B. Loftfield, Progr. in Biophys. and Biophys. Chem. **8**, 347 (1957).

[26] J. W. Littlefield, E. B. Keller, J. Gross, and P. C. Zamecnik, J. Biol. Chem. **217**, 111 (1955).

[27] T. O. Caspersson, *Cell Growth and Cell Function* (W. W. Norton and Company, Inc., New York, 1950).

[28] J. Brachet in *The Nucleic Acids*, E. Chargaff and J. N. Davidson, editors (Academic Press, Inc., New York, 1955), Vol. II, p. 475.

[29] S. Spiegelman, H. O. Halvorson, and R. Ben-Ishai, *A Symposium on Amino Acid Metabolism* (The Johns Hopkins Press, Baltimore, Maryland, 1955), p. 124.

[30] G. Schmidt, K. Seraidarian, L. M. Greenbaum, M. D. Hickey, and S. J. Thannhauser, Biochim. et Biophys. Acta **20**, 135 (1956).

[31] L. E. Hokin and M. R. Hokin, Biochim. et Biophys. Acta **13**, 236, 401 (1954).

[32] H. Borsook, Advances in Protein Chem. **8**, 127 (1953).

[33] H. Wasteneys and H. Borsook, Physiol. Rev. **10**, 110 (1930).

[34] K. U. Linderstrøm-Lang in *Lane Medical Lectures, 1951* (Stanford University Press, Stanford, 1952), Vol. VI.

[35] J. S. Fruton in *Essays in Biochemistry* (John Wiley and Sons, Inc., New York, 1956), p. 106.

[36] J. S. Fruton in *Harvey Lectures* (Academic Press, Inc., New York, 1955–1956), Vol. LI, p. 64.

[37] C. S. Hanes, F. J. R. Hird, and F. A. Isherwood, Biochem. J. **51**, 25 (1952).

[38] W. J. Williams and C. B. Thorne, J. Biol. Chem. **210**, 203 (1954).

[39] W. J. Williams and C. B. Thorne, J. Biol. Chem. **211**, 631 (1954).

[40] W. J. Williams and C. B. Thorne, J. Biol. Chem. **212**, 427 (1955).

[41] C. Zioudrou, S. Fujii, and J. S. Fruton, Proc. Natl. Acad. Sci. U. S. **44**, 439 (1958).

[42] H. Tarver in *The Proteins*, H. Neurath and K. Bailey, editors (Academic Press, Inc., New York, 1954), Vol. II, p. 1199.

[43] H. Borsook, C. L. Deasy, A. J. Haagen-Smit, G. Keighley, and P. H. Lowy, J. Biol. Chem. **186**, 297 (1950).

[44] H. Borsook, C. L. Deasy, A. J. Haagen-Smit, G. Keighley, and P. H. Lowy, J. Biol. Chem. **196**, 669 (1952).

[45] P. Siekevitz, J. Biol. Chem. **195**, 549 (1952).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

[46] P. C. Zamecnik and E. B. Keller, J. Biol. Chem. **209**, 337 (1954).
[47] E. B. Keller and P. C. Zamecnik, J. Biol. Chem. **221**, 45 (1956).
[48] J. W. Littlefield and E. B. Keller, J. Biol. Chem. **224**, 13 (1957).
[49] A. E. Mirsky, S. Osawa, and V. G. Allfrey, Cold Spring Harbor Symposia Quant. Biol. **21**, 49 (1956).
[50] V. G. Allfrey, A. E. Mirsky, and S. Osawa, J. Gen. Physiol. **40**, 451 (1957).
[51] V. G. Allfrey and A. E. Mirsky, Proc. Natl. Acad. Sci. U. S. **43**, 589 (1957).
[52] G. C. Webster, J. Biol. Chem. **229**, 535 (1957).
[53] E. F. Gale and J. P. Folkes, Biochem. J. **55**, 721 (1953).
[54] E. F. Gale and J. P. Folkes, Biochem. J. **59**, 661, 675 (1955).
[55] E. F. Gale, Biochem. Soc. Symposia **14**, 47 (1957).
[56] M. V. Simpson and J. R. McLean, Biochim. et Biophys. Acta **18**, 573 (1955).
[57] J. R. McLean, G. L. Cohen, and M. V. Simpson, Federation Proc. **15**, 312 (1956).
[58] H. M. Bates, V. M. Craddock, and M. V. Simpson, J. Am. Chem. Soc. **80**, 1000 (1958).
[59] M. Beljanski and S. Ochoa, Proc. Natl. Acad. Sci. U. S. **44**, 494 (1958).
[60] J. Mandelstam and H. J. Rogers, Nature **181**, 956 (1958).
[61] J. E. Snoke and K. Bloch in *Glutathione*, S. P. Colowick, editor (Academic Press, Inc., New York, 1954), p. 129.
[62] J. E. Snoke and K. Bloch, J. Biol. Chem. **213**, 825 (1955).
[63] J. F. Speck, J. Biol. Chem. **168**, 403 (1947).
[64] J. F. Speck, J. Biol. Chem. **179**, 1387, 1405 (1949).
[65] W. H. Elliott, Biochem. J. **49**, 106 (1951).
[66] W. H. Elliott, J. Biol. Chem. **201**, 661 (1953).
[67] L. Levintow and A. Meister, J. Am. Chem. Soc. **75**, 3039 (1953).
[68] L. Levintow, A. Meister, G. H. Hogeboom, and E. L. Kuff, J. Am. Chem. Soc. **77**, 5304 (1955).
[69] L. Levintow and A. Meister, J. Biol. Chem. **209**, 265 (1954).
[70] A. Meister, Physiol. Rev. **36**, 103 (1956).
[71] P. D. Boyer, O. J. Koeppe, and W. W. Luchsinger, J. Am. Chem. Soc. **78**, 356 (1956).
[72] A. Kowalsky, C. Wyttenbach, L. Langer, and D. E. Koshland, Jr., J. Biol. Chem. **219**, 719 (1956).
[73] H. Waelsch, Advances in Enzymol. **13**, 237 (1952).
[74] H. Waelsch, Advances in Protein Chem. **6**, 299 (1951).
[75] E. P. Abraham, *Biochemistry of Some Peptide and Steroid Antibiotics* (John Wiley and Sons, Inc., New York, 1957).
[76] M. Bergmann, L. Zervas, and L. Salzmann, Ber. dent. Chem. Geo. **66**, 1288 (1933).
[77] W. K. Maas, 3rd Intern. Congress Biochem. Abstracts 4–36, Brussels (1955).
[78] P. Berg, J. Am. Chem. Soc. **77**, 3163 (1955).
[79] P. Berg, J. Biol. Chem. **222**, 991 (1956).
[80] W. P. Jencks and F. Lipmann, J. Biol. Chem. **225**, 207 (1957).
[81] H. Chantrenne, J. Biol. Chem. **189**, 227 (1951).
[82] D. Schachter and J. V. Taggart, J. Biol. Chem. **203**, 925 (1953).
[83] D. Schachter, and J. V. Taggart, J. Biol. Chem. **208**, 263 (1954).
[84] D. Schachter and J. V. Taggart, J. Biol. Chem. **211**, 271 (1954).
[85] K. Moldave and A. Meister, Biochim. et Biophys. Acta **25**, 434 (1957).
[86] K. Moldave and A. Meister, J. Biol. Chem. **229**, 463 (1957).
[87] M. B. Hoagland, Biochim. et Biophys. Acta **16**, 288 (1955).
[88] M. B. Hoagland, E. B. Keller, and P. C. Zamecnik, J. Biol. Chem. **218**, 345 (1956).
[89] P. Berg, J. Biol. Chem. **222**, 1025 (1956).
[90] J. A. DeMoss and G. D. Novelli, Biochim. et Biophys. Acta **18**, 592 (1955).
[91] J. A. DeMoss and G. D. Novelli, Biochim. et Biophys. Acta **22**, 49 (1956).
[92] R. S. Schweet, R. W. Holley, and E. Allen, Arch. Biochem. Biophys. **71**, 311 (1957).
[93] F. Lipmann, Proc. Natl. Acad. Sci. U. S. **44**, 67 (1958).
[94] J. A. DeMoss, S. M. Genuth, and G. D. Novelli, Proc. Natl. Acad. Sci. U. S. **42**, 325 (1956).
[95] M. B. Hoagland, P. C. Zamecnik, N. Sharon, F. Lipmann, M. P. Stulberg, and P. D. Boyer, Biochim. et Biophys. Acta **26**, 215 (1957).
[96] M. Karasek, P. Castelfranco, P. R. Krishnaswamy, and A. Meister, J. Am. Chem. Soc. **80**, 2335 (1958).
[97] E. W. Davie, V. V. Koningsberger, and F. Lipmann, Arch. Biochem. Biophys. **65**, 21 (1956).
[98] G. D. Novelli, Proc. Natl. Acad. Sci. U. S. **44**, 86 (1958).
[99] P. Berg, J. Biol. Chem. **233**, 601 (1958).
[100] P. Castelfranco, K. Moldave, and A. Meister, J. Am. Chem. Soc. **80**, 2335 (1958).
[101] R. S. Schweet, F. C. Bovard, E. Allen, and E. Glassman, Proc. Natl. Acad. Sci. U. S. **44**, 173 (1958).
[102] P. Berg and E. J. Ofengand, Proc. Natl. Acad. Sci. U. S. **44**, 78 (1958).
[103] T. Hultin and G. Beskow, Exptl. Cell. Research **11**, 664 (1956).
[104] M. B. Hoagland, P. C. Zamecnik, and M. L. Stephenson, Biochim. et Biophys. Acta **24**, 215 (1957).
[105] M. B. Hoagland, M. L. Stephenson, J. F. Scott, L. I. Hecht, and P. C. Zamecnik, J. Biol. Chem. **231**, 241 (1958).
[106] K. Ogata and H. Nohara, Biochim. et Biophys. Acta **25**, 659 (1957).
[107] S. B. Weiss, G. Acs, and F. Lipmann, Proc. Natl. Acad. Sci. U. S. **44**, 189 (1958).
[108] P. Zamecnik, M. L. Stephenson, and L. I. Hecht, Proc. Natl. Acad. Sci. U. S. **44**, 73 (1958).
[109] R. W. Holley, J. Am. Chem. Soc. **79**, 658 (1957).
[110] J. L. Potter and A. L. Dounce, J. Am. Chem. Soc. **78**, 3078 (1956).
[111] L. I. Hecht, M. L. Stephenson, and P. C. Zamecnik, Federation Proc. **17**, 239 (1958).
[112] J. H. Northrup, M. Kunitz, and R. M. Herriott, *Crystalline Enzymes* (Columbia University Press, New York, 1955), second edition, p. 232.
[113] M. M. Daly and A. E. Mirsky, J. Gen. Physiol. **36**, 243 (1952–1953).
[114] P. N. Campbell and T. S. Work, Biochem. J. **52**, 217 (1952).
[115] H. Green and H. S. Anker, J. Gen. Physiol. **38**, 283 (1955).
[116] F. W. Putnam, F. Meyer, and A. Miyake, J. Biol. Chem. **221**, 517 (1956).
[117] T. Peters, Jr., J. Biol. Chem. **200**, 461 (1953).
[118] T. Peters, Jr., Federation Proc. **16**, 369 (1957).
[119] J. L. Simkin and T. S. Work, Biochem. J. **67**, 617 (1957).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

26
Coding and Information Theory

PETER ELIAS

Department of Electrical Engineering and Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge 39, Massachusetts*

INTRODUCTION

"INFORMATION theory" is used in at least three senses. In the narrowest of these senses, it denotes a class of problems concerning the generation, storage, transmission, and processing of information, in which a particular measure of information is used. This area is also called "coding theory" and, especially in Britain, "the mathematical theory of communication," which is the title of Shannon's original paper[1] from which the field is derived. This is the usage in the titles of the books by Khinchin[2] and Feinstein,[3] which are rather abstract mathematical presentations.

In a broader sense, information theory has been taken to include any analysis of communications problems, including statistical problems of the detection of signals in the presence of noise, that make no use of an information measure. Woodward[4] has shown the relationship of information measure to some of these problems. A book by Wiener,[5] an article by Rice[6,7] (reprinted in a book edited by Wax[8]), and a recent textbook by Davenport and Root[9] discuss the use of statistical techniques in problems concerning analysis of signals and noise, and books by Blanc-LaPierre and Fortet,[10] Loeve,[11] and Doob[12] provide the (abstract) mathematical background. It is in this broader sense that the word is used, for example, in the title of the Professional Group on Information Theory of the Institute of Radio Engineers, whose *Transactions* contains articles both on coding and on signal-noise problems.

In a still broader sense, information theory is used as a synonym for the term "cybernetics" introduced by Wiener[13] to denote, in addition to the areas listed in the foregoing, the theory of servomechanisms, the theory of automata, and the application of these and related disciplines to the study of communication, control, and other kinds of behavior in organisms and machines. This is the usage in the titles of three meetings held in London (two of which have published proceedings[14,15]) and two held at the Massachusetts Institute of Technology.[16,17] *Information and Control* publishes articles in this broad area.

Only the area covered by the narrowest definition is discussed here—not because it is necessarily the most important for biological applications, but because of the limitations of space.

* The work of this Laboratory is supported in part by the U. S. Army (Signal Corps), the U. S. Air Force (Office of Scientific Research, Air Research and Development Command), and the U. S. Navy (Office of Naval Research).

INFORMATION MEASURE. SOURCES

The first problem is to assign a measure to information. Figure 1 shows two representations of a message and the code book which connects them. At first one might suppose that a long message has more information than a short one, but, as the figure shows, the message length—or any other characteristic of the representation itself—can be changed drastically by coding at the transmitter. If the receiver decodes correctly, the message has been transmitted successfully by using a very brief form. In a communications system, any message has a variety of different representations in different places. One may want to say that it contains the same amount of information. An amount of information, then, cannot depend upon the form of representation. It seems reasonable, however, to make it depend upon how many messages there are. If the number of messages is large, longer code words will have to be used in order to distinguish between them. One starts, therefore, with the hypothesis that the information in a message is some function of the number of messages in the set. One would like to say that two successive, independent selections from the same set have twice the information value of a single selection, *if the two messages are equally probable.* This demands a function $f(m)$ of the number m of messages in the set, for which $f(m^2)=2f(m)$, and the logarithm is the only respectable function with this property. For a selection between two messages, this gives $[\log 2]$ units of information. This can be rewritten as $[-\log(\frac{1}{2})]$ or, in general, as the negative of the logarithm of the message probability. This, in fact, is what one chooses to gener-

CODED MESSAGES

ABBA . . .

THE QUICK BROWN FOX JUMPED OVER THE LAZY DOG NOW IS THE TIME FOR ALL GOOD MEN TO COME TO THE AID OF THE PARTY NOW IS THE TIME FOR ALL GOOD MEN TO COME TO THE AID OF THE PARTY THE QUICK BROWN FOX JUMPED OVER THE LAZY DOG . . .

CODE BOOK

A↔THE QUICK BROWN FOX JUMPED OVER THE LAZY DOG

B↔NOW IS THE TIME FOR ALL GOOD MEN TO COME TO THE AID OF THE PARTY

I(A)=I(THE QUICK BROWN FOX JUMPED OVER THE LAZY DOG)=LOG 2=−LOG(1/2)=−LOG Pr[A]=1 BIT/SYMBOL

FIG. 1. Two representations of a message and the code book which connects them.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

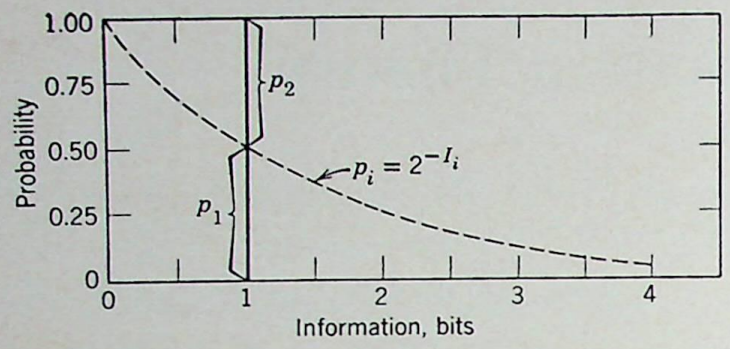


FIG. 2. Distribution of information for a two-symbol source whose symbols have equal probabilities.

$p_1 = p_2 = \frac{1}{2}, \quad I_1 = I_2 = -\log\frac{1}{2} = 1$ bit

$\bar{I} = H = \frac{1}{2} + \frac{1}{2} = 1$ bit

alize as the information associated with a message. The choice of a logarithmic base determines the unit of information, and the base 2 is chosen, the unit being the "bit." Thus, a selection between two equiprobable alternatives requires one bit of information.

This definition is plausible, but it does not justify the information measure. The adopted measure is justified, in a practical sense, on the average, because a source that generates information more rapidly than another, also requires more communications facilities—more band width, time, signal-to-noise ratio—to transmit its output successfully.

The information source, Fig. 1, is characterized by the distribution of information, Fig. 2. Since this source generates one bit of information per selection regardless of which message is selected, its information distribution is degenerate: all of the probability is piled up at one bit. Each symbol has a probability given by the value of the exponential at the information value of the symbol, as the plot is just probability *vs* minus log probability. Figure 3 shows the information distribution for another two-symbol source whose symbols have the unequal probabilities $\frac{3}{4}$ and $\frac{1}{4}$. This source generates only 0.42 bits when it selects an *A* and 2.0 bits when it selects a *B*. Its average rate is 0.81 per symbol, which is smaller than in Fig. 2 with equiprobable symbols. This is a general characteristic of sources. Any constraint on the number of symbols or sequences that a source may generate and any shift away from equal probabilities will reduce the average source rate. A source constrained to generate sequences of letters which spell out sentences in English has a lower average rate than a source which selects successive letters of the alphabet with statistical independence.

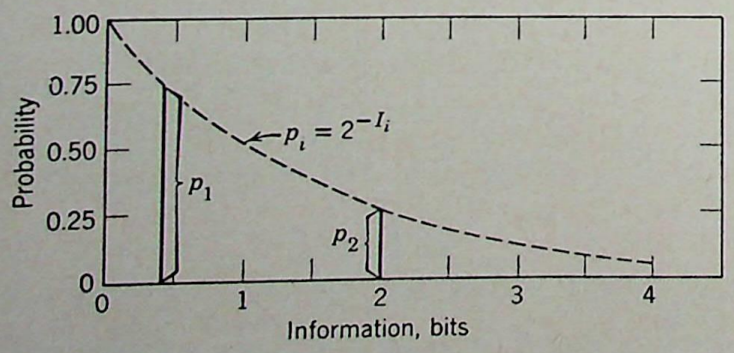


FIG. 3. Distribution of information for a two-symbol source whose symbols have the unequal probabilities $\frac{3}{4}$ and $\frac{1}{4}$.

$p_1 = \frac{3}{4}, \quad I_1 = -\log\frac{3}{4} = 0.42$ bit

$p_2 = \frac{1}{4}, \quad I_2 = \log\frac{1}{4} = 2.00$ bits

$\bar{I} = \sum_{i=1}^{2} p_i I_i = H = \sum_{i=1}^{2} -p_i \log p_i = 0.31 + 0.50 = 0.81$ bits

There are, of course, many questions which arise in connection with coding and with the description of sources having sequential constraints. What has been covered here is a mathematical theory, and problems arise in its application. One of these is illustrated in Fig. 4. The problem consists in identifying the alphabet which is relevant in a given situation. Suppose that, in observing the output of a neuron, the eight wave forms shown in the figure are seen with equal frequency. This would give three bits of information per wave form. However, these signals may not all be distinguishable by the system being observed. The system may act only as a pulse counter over this time interval, and may recognize only four different signals—no pulse,

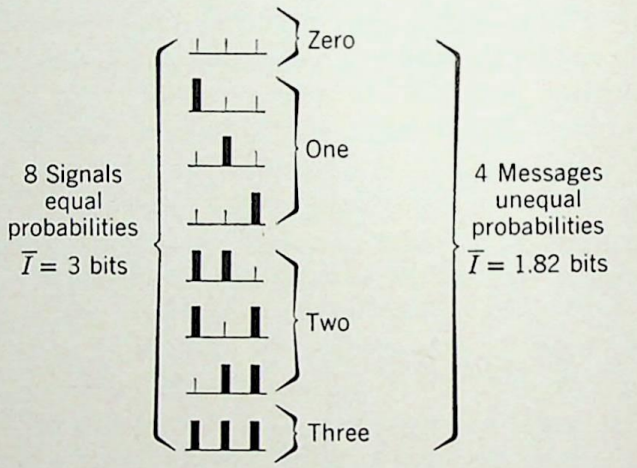


FIG. 4. Dependence of source rate on alphabet.

one pulse, two pulses, and three pulses. Then the average rate is only 1.82 bits per signal. There is no way of telling from the outside which of these alphabets is actually in use—if, in fact, either is. It is necessary to determine that the system does have different responses to two signals before they can be defined as being distinct letters of an alphabet. The problem of recognizing the relevant alphabet always is present and shows up in many ways, including the selection of scales of resolution to be used in amplitude and time to distinguish different signals. In Fig. 4, the average rates for the two alphabets are not too far apart, but in a train of 100 pulses, if all observed wave forms are equiprobable, there is a factor of about 20 between the rates obtained for the two corresponding alphabets.

NOISY CHANNELS

There are, then, sources to select symbols, and channels are needed to transmit them. Figure 5 shows one—a noisy channel. In communications, a noisy chan-

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

nel usually is a medium which separates transmitter and receiver. In information storage, a noisy channel may model the action of the environment which, through thermal agitation or other forces, may cause changes in the stored information. In Fig. 5, the channel eliminates some of the distinctions present in its input. The source selects from among four equiprobable symbols a, b, c, and d. The receiver on reception of A or B can eliminate two input possibilities, but it cannot choose between the other two.

To analyze informationally what happens in the channel, consider a particular transmission event: the selection of a at the transmitter and the reception of A at the receiver. It is assumed that the receiver knows the input-letter probabilities and the channel. Thus, before reception of A, the estimate by the receiver of the probability that a will be transmitted is $\frac{1}{4}$. After reception of A, the receiver knows that only either a or b could have been transmitted, and both are equally likely. Thus, *a posteriori*, the probability of a (on the evidence A) is $\frac{1}{2}$.

The receiver *a priori* needs log 4 = 2 bits of information to select a as the transmitted letter. *A posteriori*, it still needs log 2 = 1 bit to select a after receiving A. The channel is credited with the 1-bit difference, which is defined as the amount of information which the receipt of A gives about the transmission of a. The quantity expressed in Fig. 5 as $I(a; A)$ is called the *mutual information* of A about a, or the *transmitted information*.

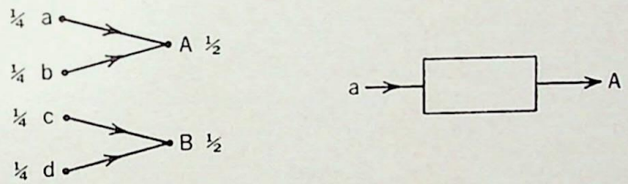


FIG. 5. The mutual information between a transmitted symbol and a received symbol.

$$Pr(a)=\tfrac{1}{4} \qquad Pr(a|A)=\tfrac{1}{2}$$
$$I(a)=2 \text{ bits} \qquad I(a|A)=1 \text{ bit}$$
$$\Delta I=\log Pr(a)-[-\log Pr(a|A)]=1 \text{ bit}$$
$$=\log\frac{Pr(a|A)}{\Pr(a)}=\log\frac{Pr(a,A)}{Pr(a)Pr(A)}=I(a;A)$$

In general, for each possible pair of transmitted and received symbols, x_i and y_j, there is a probability of occurrence $\Pr(x_i, y_j)$ [which may be zero as in Fig. 5 for $\Pr(a, B)$] and an information value $I(x_i, y_j) = \log \Pr(x_i, y_j)/\Pr(x_i)\Pr(y_j)$, which measures the change in the logarithm of the probability of x_i owing to knowledge of y_j. This quantity is positive if y_j makes x_i more probable than it was, and is negative if y_j makes x_i less probable than it was. A plot can be made of the *mutual-information distribution* of a channel-source combination. For the channel of Fig. 5, this plot is exactly like the plot of Fig. 2: one bit of information is always transmitted regardless of which input-output pair happens.

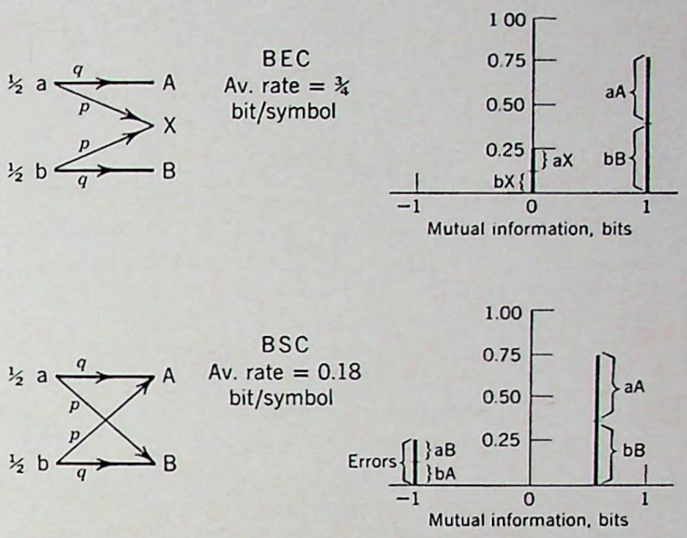


FIG. 6. The binary erasure channel, the binary symmetric channel, and their mutual information distribution.

Two more-complicated channels and their mutual-information distributions are shown in Fig. 6. In the binary erasure channel (BEC), one of two symbols is selected at the transmitter. The input symbol may be received correctly, with probability q taken here as $\frac{3}{4}$, or it may be erased with probability $p=\frac{1}{4}$: The receiver then receives an X. Evaluating mutual information gives the distribution shown. The channel sends one bit per symbol three-quarters of the time, when no erasure occurs; one-quarter of the time, when the transmitted symbol is erased, no information is transmitted. The average rate is just $\frac{3}{4}$ bit per symbol, if the input symbols are equiprobable.

In the binary symmetric channel (BSC), there are only two input and two output symbols, and true errors occur. Here less than one bit is transmitted when there is no error, and a negative amount of mutual information is transmitted if there is an error. One takes error probability $p=\frac{1}{4}$. The average rate of transmission here is only 0.18 bits per symbol with equiprobable inputs.

Returning to Fig. 5, one interprets the average rate of transmission of 1 bit per symbol as the average rate at which the receiver's ignorance is reduced—from 2 bits to 1 bit for each letter transmitted. However, the receiver's ignorance is not reduced to zero. There is an apparent difference between this channel and another with the same rate which transmits two equiprobable symbols without error. In the latter case, the ignorance of the receiver is reduced from 1 bit to zero bits for each transmission. One would like to be able to say that the average rate over a channel is the significant parameter. This requires that, for the channel of Fig. 5, one finds some method for which the rate of putting information in is reduced in such a way as to reduce both the initial and final ignorance of the receiver, keep-

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

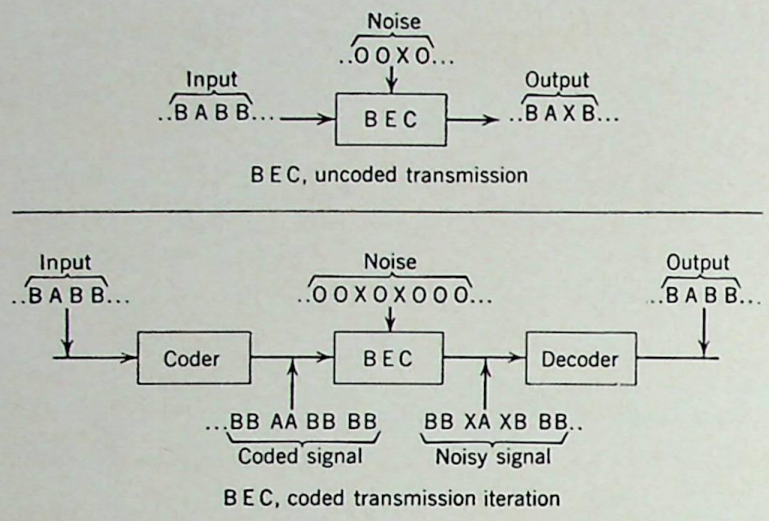


FIG. 7. Transmission over the binary erasure channel uncoded and coded by iteration.

ing their difference fixed. In this case, it is easy to do. The transmitter agrees to send only two symbols, a and c, with equal probabilities. The receiver then can decode unambiguously, receiving one bit of mutual information per symbol with no residual ignorance.

CODING FOR NOISY CHANNELS

The foregoing result is, in fact, general for a broad class of noisy channels, although the way in which the input information rate is reduced is usually more complicated. It is not possible in the BEC or BSC to find a set of input symbols which will leave the receiver with no ambiguity. All input symbols are used, therefore, but they are used in sequences, and only a fraction of the possible sequences occur.

Figure 7 shows the BEC, first used in ordinary, uncoded transmission and then used with a simple coding and decoding scheme. In the upper picture, one bit per symbol is put into the channel, $\frac{1}{4}$ of the output symbols are erased, and an average of $\frac{3}{4}$ bit per symbol comes out. Below, a coder is added, which duplicates each input symbol. For a fixed channel, input symbols can be accepted now only half as often, since two symbols go into the channel for each input symbol to the coder. The input rate is then $\frac{1}{2}$ bit per channel symbol. Again $\frac{1}{4}$ of the output symbols are erased, but both of the copies of an input symbol are erased $\frac{1}{16}$ of the time only, so reliability has improved.

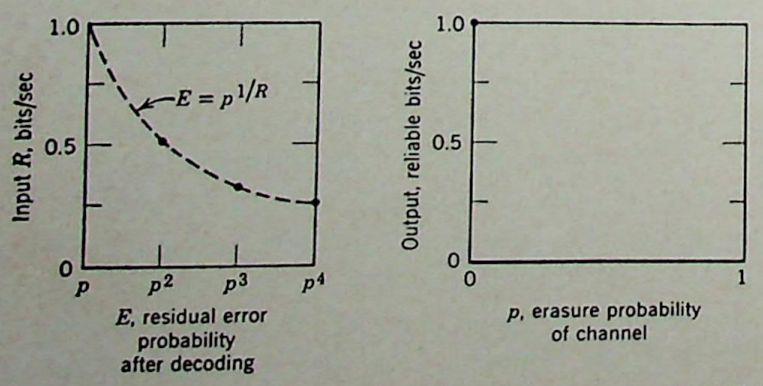
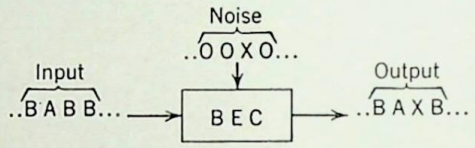


FIG. 8. Rate and error probability for iteration coding of the BEC.

Of course, each digit could be sent three or more times instead of twice. This would reduce further both the rate of transmission and the probability of total erasure, or of ambiguity after decoding. The relationship is shown in Fig. 8. On the left plot, it can be seen that sufficient reduction of rate reduces residual probability of total erasure as low as is desired, but, to get arbitrarily low probability, an arbitrarily low rate of transmission is necessary. If one looks at the channel in a different way, demanding transmission with *arbitrarily low* total-erasure probability, and asking how average rate varies as the channel-erasure probability p is varied, one sees the plot on the right in Fig. 8. If $p=0$, one can send one reliable bit per symbol, but if $p\neq 0$, none can be sent. Time must be spent on repeating the first input symbol to make its reliability arbitrarily good, and one never can get around to sending anything else.

Certainly, one way of avoiding misunderstanding is never to say anything new, but it would be discouraging



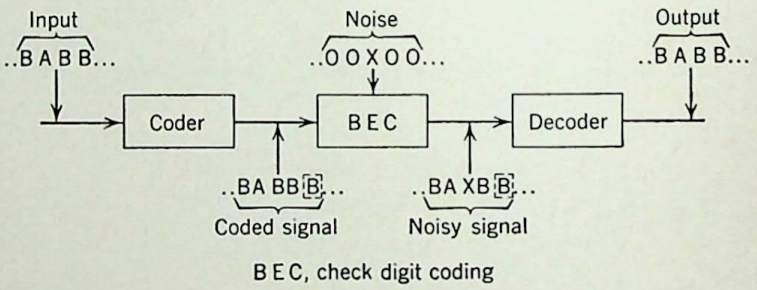


FIG. 9. The transmission erasure channel, the binary symmetric channel, and their mutual information distribution with check digits.

if this were the only way. An alternative is shown in Fig. 9.

Here one starts with uncoded transmission again. Now, however, instead of duplicating each input symbol, a general purpose replacement, called a check digit or a parity check, is inserted so that a replacement will be available in case it it erased. The digit enclosed in the dotted box is selected to make the total number of B's in the sequence of 5 symbols an *even* number. At the receiver, if only one symbol has been erased, the knowledge that an even number of B's should be present makes the decoding unique. The rate of transmission has been reduced to $\frac{4}{5}$ bit per symbol and increased reliability is obtained. There is still a danger, however, that two or more erasures may occur in the same block of five digits, in which case decoding still would be ambiguous.

The situation can be improved by adding further

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

-check digits as shown in Fig. 10. Here the sequence of symbols is shown above. Added check digits have bars over them. The check digits are computed as shown below: each symbol to the right of a row is selected to make the total number of B's in the row even; each symbol at the foot of a column is selected to make the total number of B's in the column even. First, those rows with single erasures are decoded, and then the columns having only single erasures remaining after row correction. In the case illustrated, this corrects all erasures. Further high-order check digits can be added indefinitely to give a total-erasure probability which is arbitrarily small without reducing the transmission rate to zero. Figure 11 shows the kind of relationship between rate and residual-erasure probability which results. Here an erasure probability p of 1/20 and an initial check group of ten, rather than of five, symbols have been used. Rate still goes down as reliability increases, but it now has a positive asymptote at 0.80 bits per symbol, at which rate it is possible to get arbitrary reliability. The plot of reliable rate *vs* the channel-erasure probability p is now continuous, as shown by

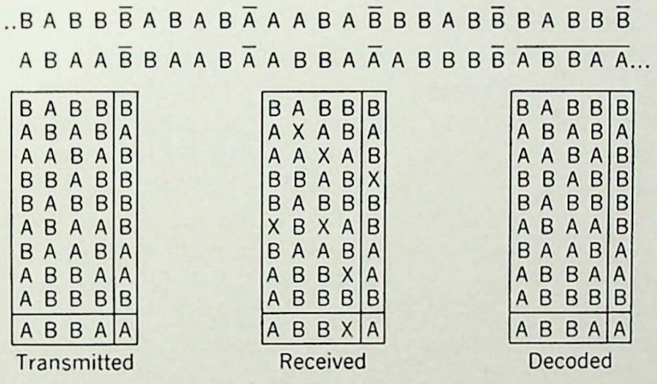


FIG. 10. Correction of erasures by iteration of parity-check digits.

the solid curve. This kind of iterated check-symbol coding can be used for the BSC[18] as well as for the BEC.[19]

Although a very special case of noisy-channel coding has been shown, it illustrates a general result. Given any noisy channel and source, with an average rate of transmission, it is possible to code the input in long blocks and to reduce both the input rate and the receiver's residual uncertainty, while keeping the average transmission rate fixed. The ideal behavior is illustrated in the dotted line on Fig. 11 for the BEC; to approach it, coding in large blocks is required.

One other point should be noted. The code we constructed was engineered carefully, and its reliability may appear to be atypical. In fact, by random selection of long sequences of symbols, it is possible to get the same results—the "ideal behavior" of Fig. 11 is obtained by just such random coding. It is necessary only to *not* select too many possible input sequences. Thus, a code suitable for reliable transmission over a noisy channel might occur quite accidentally. However, no

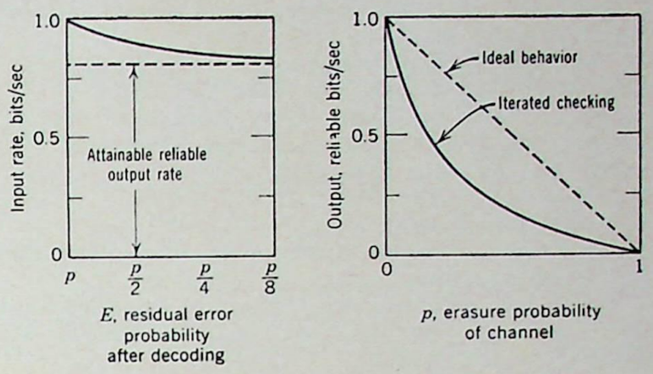


FIG. 11. Rate and error probability for iterated parity checking.

accident seems likely to account for the necessary decoding equipment and its organization.

CONCLUSION

In conclusion, it might be appropriate to point out what applications have been made to biological problems. The most successful ones, I think, have been to experiments in human communication where the channel capacity of a human being for handling information has proved a useful concept in a number of experimental situations [see e.g., Rosenblith (p. 485) and reference 20]. There also have been applications at the neurophysiological level, but in terms of statistical signal analysis rather than of information theory *per se;* some of these are referred to in the second paper by Rosenblith (p. 532). Computations of neuron channel capacity have been made, but they are of dubious value in view of the alphabet problem illustrated in Fig. 4. It seems likely that real applications are forthcoming at this level.

There have been applications to chemical specificity, etc., in biological systems (see, e.g., references 21 and 22). My feeling is that these use information measure either as a language for the discussion of purely combinatorial problems or as a useful statistic, but they do not use it in any coding sense which would imply that the informational treatment was at all necessary or unique. I think that the only other immediate application might arise in connection with the genetic-coding problem. Here, the most urgent need obviously is more data about nucleotide-amino group correspondences and the statistics of series of each. Although informational ideas may be useful here, it seems unlikely that they are essential. That is, it seems unlikely that high orders of redundancy and error-correction are being used. Some data on how local the coding is would be useful in reaching a decision on this point.

BIBLIOGRAPHY

[1] C. E. Shannon, Bell System Tech. J. **27**, 379, 623 (1948); reprinted in C. E. Shannon and W. Weaver's *The Mathematical Theory of Communication* (The University of Illinois Press, Urbana, Illinois, 1949).

[2] A. I. Khinchin, *Mathematical Foundations of Information Theory* (Dover Publications, New York, 1957).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

[3] A. Feinstein, *Foundations of Information Theory* (McGraw-Hill Book Company, Inc., New York, 1958).

[4] P. M. Woodward, *Probability and Information Theory* (McGraw-Hill Book Company, Inc., New York, 1953).

[5] N. Wiener, *The Extrapolation, Interpolation and Smoothing of Time Series* (Technology Press, Cambridge, Massachusetts and John Wiley and Sons, Inc., New York, 1949).

[6] S. O. Rice, Bell System Tech. J. **23**, 282 (1944).

[7] S. O. Rice Bell System Tech J. **24**, 46 (1945).

[8] N. Wax, editor, *Noise and Stochastic Processes* (Dover Publications, New York, 1954).

[9] W. B. Davenport, Jr., and W. J. Root, *An Introduction to the Theory of Random Signals and Noise* (McGraw-Hill Book Company, Inc., New York, 1958).

[10] A. Blanc-LaPierre and R. Fortet, *Théorie des fonctions aléatoires* (Masson et Cie., Paris, 1953).

[11] M. Loeve, *Probability Theory* (D. Van Nostrand Company, Inc., New York, 1955).

[12] J. L. Doob, *Stochastic Processes* (John Wiley and Sons, Inc., New York, 1953).

[13] N. Wiener, *Cybernetics* (Technology Press, Cambridge, Massachusetts and John Wiley and Sons, Inc., New York, 1948).

[14] W. Jackson, editor, *Proceedings of a Symposium on Application of Information Theory, London, 1952* (Butterworth Scientific Publications, London, 1953).

[15] C. Cherry, editor, *Information Theory, Proceedings of a Symposium* (Butterworth Scientific Publications, London, 1956).

[16] IRE Trans. Professional Group on Information Theory **PGIT-4** (1954).

[17] IRE Trans. on Information Theory **IT-2** No. 3 (1956).

[18] P. Elias, IRE Trans. Professional Group on Information Theory **PGIT-4**, 29 (1954).

[19] P. Elias, in *Handbook of Automation, Computation and Control*, E. Grabbe, editor (John Wiley and Sons, Inc., New York, 1958), Vol. I, p. 16–01.

[20] H. Quastler, editor, *Information Theory and Psychology* (Free Press, Glencoe, Illinois, 1955).

[21] H. Quastler, editor, *Essays on the Use of Information Theory and Biology* (The University of Illinois Press, Urbana, Illinois, 1953).

[22] H. P. Yockey, R. L. Platzman, and H. Quastler, editors, *Symposium on Information Theory in Biology* (Pergamon Press, New York, 1958).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

REVIEWS OF MODERN PHYSICS VOLUME 31, NUMBER 1 JANUARY, 1959

27
Genetics and the Role of Nucleic Acids

CYRUS LEVINTHAL

Department of Biology, Massachusetts Institute of Technology, Cambridge 39, Massachusetts

THE purpose of this article is to review some of the basic ideas in genetics and to set the stage for a discussion of certain specialized problems. The first question to be considered is the nature of the genetic map: its construction and meaning. The genetic map is an abstract entity defined in terms of the probabilities with which certain types of offspring are produced in matings. In a sense, it is related to a physical structure of the cell; however, this relationship must be established experimentally, and it is not a logical consequence of the definition of the map.

For any organism which can duplicate without a sexual process, one finds that progeny organisms are, in general, identical with the parental. Occasionally, an alteration occurs. If this alteration persists—that is, if all of the descendants of the altered organism have the same alteration—then, the change is called a mutation. In an operational sense, mutations form the basic entities of genetics. Mutations may occur which lead to red hair, to six toes, or to any of a number of other attributes in a higher organism. Or for bacteria, a mutation may result in a small colony or in a colony which is red in a certain medium. For the moment, assume that one can deal with several such mutations which are recognizable and recognizably different from one another. Furthermore, assume that an organism which contains several such mutations can be detected, and that the total number and types of mutations it contains can be enumerated. The discussion of the possible physicochemical bases of such mutations is undertaken in a later article (p. 249). For the moment, it is only necessary that the mutations give a recognizable result. The analysis begins with the definition of some organism as the standard or "wild" type. Suppose that one has three different, recognizable mutations which arise from the wild type; call them A, B, and C. The standard wild-type organism is called, by definition, "plus" with respect to each of these characters. The organism with the first mutation having lost the property A, is designated $A\ B+\ C+$; that is, it has lost the property A, but maintains the properties $B+$ and $C+$. Organisms of this type also may be designated simply by the letter "A".

A group of organisms may be selected, one of which has the mutation A, one B, and one C. If one works with microorganisms, then from a single individual of each of these types, stocks containing many millions of microorganisms can be grown. Except for secondary mutations, each organism within a stock will have lost a specific property.

Any two of these mutant types may be allowed to mate with one another. The mechanism of the mating is not of concern at the moment. It will suffice to say that, in general, the experiments with microorganisms involve mixing of the two types under suitable conditions and allowing the interactions to occur at random.[1] If a stock of organisms of one type A is allowed to mate with a stock of another type B, the experiment is called a cross between A and B and is described by the symbol $A\times B$.

The organisms discussed here are called haploid, meaning that they have only one set of genetic determinants. In many higher organisms, complications can arise which result from the double set of genetic determinants which they contain. With regard to the discussion of a genetic map, none of these complications represents an alteration to the general conceptual scheme. Thus, for the purpose of simplification, they are left out entirely and only haploid organisms are considered.

Among the progeny arising from the cross $A\times B$ are found organisms of the type A and organisms of the type B which are identical with either of the parents. Progeny organisms also are found which lack both the character A and the character B. Still others are found which are like the wild type in having both $A+$ and $B+$. A and B are parental in their genetic composition and the two types $A\ B$ have the property B (i.e., the loss of $B+$) and the property A (i.e., the loss of $A+$) which it must have received, one from one parent and one from the other parent. And the organism $A+B+$ must have received the property $A+$ from one parent and the property $B+$ from the other parent. The new types, $A\,B$ and $A+B+$, are called *recombinants*.

The fraction of the descendants which are of the recombinant types is called the recombinant frequency and is the only number which is needed to describe the results of the cross. This is true, however, only if there has been no selection of one of the types in preference to the others. If a cross has been carried out involving equal numbers of A and B parents and if the analysis takes into account any possible difference in the mating type (maleness or femaleness), then one finds that the two recombinant types $A\ B$ and $A+B+$ occur with equal frequency. Again, this is true *only* if the presence of the property $A+$ or the property $B+$ in the organism does not give it a selective advantage in the growth which occurs before the final progeny are selected and scored. If there were any such selective advantage in the growth, it would constitute a physiological effect which is not relevant to the genetic phenomenon being discussed. Therefore, only those cases are considered in

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

which the reciprocal recombinants arising from a cross are, on the average, equal to each other. In practice, this is not a serious limitation, since in most cases such selection easily is eliminated.

With a given pair of mutations, two crosses can be carried out. The two mutant types can be crossed with each other; that is, $A\ B+ \times A+B$, or the double mutant $A\ B$ can be crossed with the wild type $A+B+$. In the first case, the recombinants are $A\ B$ and $A+B+$ while in the second they are $A\ B+$ and $A+B$. For the systems discussed here in which none of the possible types has any growth advantage over the others, the resulting recombinant frequency found in these two different crosses is the same. Thus, one may conclude that there is only one number needed to describe the results of crosses between any pair of mutants—namely, the recombinant frequency in the progeny.

In dealing with three different mutants, three different crosses can be carried out between pairs of them: A with B, A with C, and B with C. The recombinant frequency obtained in $A \times B$ is designated $A \cdot B$, and the same notation is used for the recombinant frequency in $B \times C$ and $A \times C$. If the markers are neither too close nor too distant from each other, it is found that the following relationship holds as a good first approximation:

$$A \cdot B + B \cdot C = A \cdot C.$$

That is, two of the three recombination frequencies, when added together, equal a third. This additivity obtains if the recombination frequencies are not too large. If the recombination frequency is large, since recombination can be assumed to result from some kind of switching event, then there is a high probability of more than one switch taking place in the same mating. If an even number of switches occur between two genetic markers, then no recombinant is observed. Correcting for these multiple events by calculating the total probability of switches, under the assumption that they occur at random, one finds that the results of most crosses are now consistent with the additivity relationship except when extremely closely linked markers are involved. To the extent that this relationship holds, mutants can be arranged as points along a line, the distances between the points being proportional to the recombination frequencies. This line with the mutants arranged along it is called the *genetic map*.

The genetic map defined in this way has two important characteristics which must be distinguished. The first is that the mutants can be arranged in unique order along a line without requiring any branch points or any other configuration which implies a two-dimensional surface. The second is that the distances along the line are proportional to the recombination frequencies. For the latter to be true, strict additivity must hold. If it does hold, then it is true also that a one-dimensional map will be demonstrated.

However, the one-dimensional character of the map can be demonstrated even if additivity does not apply. Three mutations can be ordered with respect to each other in two different ways. As in the foregoing, one can measure the recombination frequencies when each of the three pair-wise crosses are performed. The pair which shows the highest recombination frequency between them can be taken as the outside pair, while the one which shows a lower recombination frequency with each of the other two can be placed between them. A more precise method of ordering three mutations is by means of a three-factor cross. This method utilizes the fact that if the probability is low for a switching event which produces a recombination, the occurrence of two such events is much less likely than the occurrence of a single one. Three-factor crosses are done if an organism has, for example, both mutations A and B crossed with an organism which has mutation C. The other two crosses of this type also can be done—namely, $A\ C \times B$ and $B\ C \times A$. In these crosses, one can examine the progeny for recombinants which are like the wild type in that they contain all three properties, $A+$, $B+$, and $C+$. One of the three crosses will yield, in general, fewer recombinants of this type than the other two. Suppose, for example, the cross $A\ C \times B$ produces the smallest number of recombinants. This would imply that A and C are on the outside and B between them.

This ordering of three elements in an operationally unique way still implies nothing about the nature of the genetic map, since the ordering, by definition, is always possible. But suppose that there is a large number of mutations and all possible triplets are ordered in the same way. Then, it is possible to test whether or not the data are consistent with a one-dimensional map. In all cases so far studied, it has been found that the mutants can be arranged uniquely along one or more linear elements. In some cases, one set of mutants will assort randomly with respect to another set of mutants in crosses. For extremely close markers (those which give very little recombination in crosses between them), one finds a marked deviation from additivity. This is related to the phenomenon of interference in which the occurrence of one switch has an influence on the probability of another occurring immediately adjacent to it. In some systems, the probability of the second switch is reduced, in which case the effect is called interference. In other systems, the probability of multiple switches in the same region is very high, in which case the effect is called negative interference. In the instance of bacteriophage, strong negative interference is found,[2] only for sets of mutants which are very closely linked. Benzer[3] has developed an extremely elegant and essentially new method of establishing the one-dimensionality of the genetic map by making use of mutants which act as though they extend over a finite length of the genetic map. Thus, the conclusion from many types of genetic analyses is that independently arising muta-

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

tions can be ordered along one-dimensional structures whether strict additivity applies or not.

So far, the recombinational event itself has been discussed in very generalized terms. It is an event which occurs with constant probability per unit length along the genetic map and which results in the formation of recombinants. For higher organisms, there is good reason to believe that recombinants are formed by *crossing over*. In this process, the two homologous structures which correspond physically to the genetic map pair each with the other, with a break occurring at the corresponding points. The broken ends reunite in such a way as to form two new structures, each of which contains parts of the genetic structure of the parents. One of the characteristics of the crossing-over model is that recombinants always are formed in reciprocal pairs; that is, if the cross $A \times B$ is done, the wild-type and the type $A\ B$ recombinants always are formed in the same event. This is a much stronger statement than that made earlier about the equality of these reciprocal recombinants on the average.

Only by special tricks can one determine whether or not reciprocal recombinants are formed in the same act. However, in those higher organisms where the phenomenon can be studied—and this is primarily with the molds, which, for present purposes, are considered as "higher"—one can show that the predictions of the crossing-over model are borne out. Where the corresponding experiments are possible with viruses and bacteria, it is generally found that reciprocal recombinants are not formed in the same event. In those cases where recombinants are formed without associated formation at their reciprocals, there is also evidence which suggests that the parental structures are not broken during the recombination act. One is then led to a model of recombinant formation which involves the copying of the genetic information along one structure for part of its length, followed by a shift after which the copying continues along another parent. This mechanism has been called *copy-choice* (Fig. 1).

It should be kept in mind that the generalizations about the genetic map are independent of the mechanism of recombination. Whether recombinants arise by crossing-over or by copy-choice, the one dimensionality of the map and the roughly additive distances along it can be established. The event which produces the recombinants, regardless of its nature, must, by definition, occur with a constant probability per unit length of the map.

The next question concerns the relationship between the genetic map and the physical constituents of the cell. In the case of a few higher organisms, it has been possible to show by very direct methods that there is a one-to-one correspondence between the position of a genetic marker along the map and the position of an observable alteration on the chromosome of the organism.[4] The chromosomes are cellular components origi-

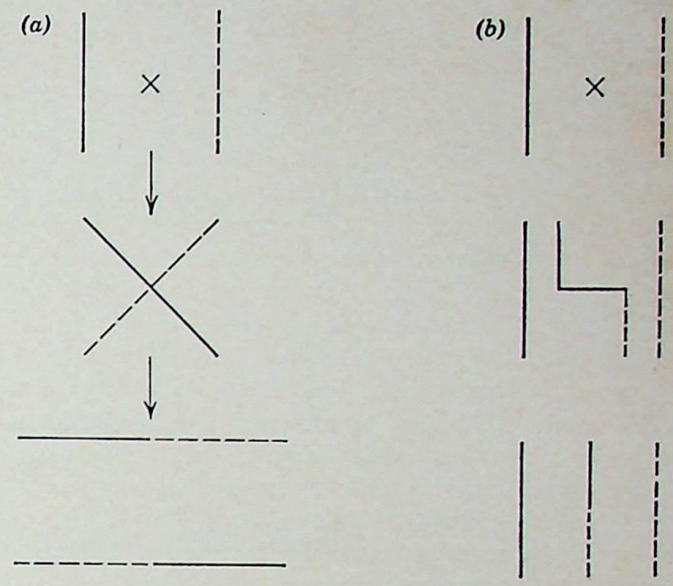


FIG. 1. Two models by which recombinants could be formed considering the genetic map only; (a) is the crossing-over model and (b) is copy-choice.

nally characterized by the color they demonstrate with particular stains. The chromosomes of the *Drosophila* salivary gland are abnormally large and, because of characteristic banding, individual portions can be identified by direct microscopic observation. Inverted or missing segments of a chromosome arise in some mutant organisms and can be recognized morphologically. Corresponding to the inversion in the visible chromosome is an inversion in the mapping of a short region of the genetic map. The same holds true when deletions are observed cytologically. In this case, a small region of the genetic map acts as though it had been deleted. By experiments with mutants of this type, it has been possible to establish a one-to-one correspondence between the position along the genetic map and the position along a chromosome (Fig. 2). This correspondence, however, has been made in detail only with the giant chromosomes of *Drosophila* which contain approximately a thousandfold more material than normal chromosomes of *Drosophila* and are in cells which cannot undergo further division.

In all of the organisms where it is technically possible to do the combined cytological and genetic studies, the correspondence between the genetic map and the chromosomes can be demonstrated. The chromosome, however, is a complicated structure which contains several chemical components. Of these, only deoxyribonucleic acid (DNA) is constant in amount in all normal cells of a given organism. Those cells which have only one set of genetic determinants contain only half as much DNA as the amount contained in the majority of cells known to possess a double set of chromosomes.

There is a considerable body of evidence indicating that the chemical structure which contains the genetic information and which, in fact, corresponds to the

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

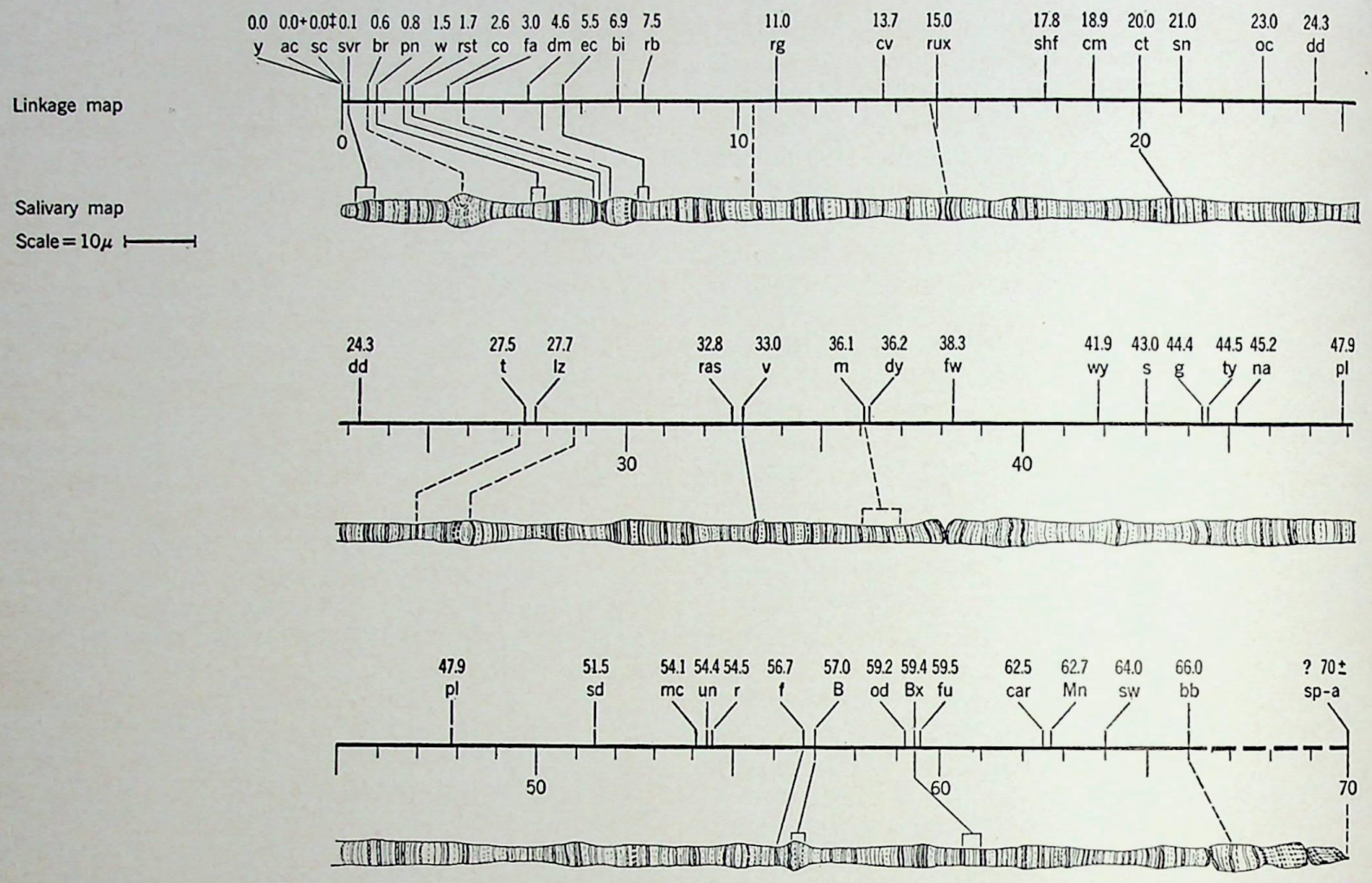


FIG. 2. The genetic map (upper line) and the *X*-chromosome of the salivary gland of *Drosophila Melanogaster* [after J. L. Bridges, J. Heredity **29**, Suppl. 1 (1938)].

genetic map is nucleic acid. Since this conclusion has come to be very widely accepted, it is important to examine in some detail the experimental basis for it. Four types of experiments are reviewed which lend strong support to the view that nucleic acid can carry genetic information. The acceptance of the generalization, however, that nucleic acids do carry the genetic information in all organisms is still very much a matter of personal taste. Most of the direct experimental demonstrations have come from studies with microorganisms, and, in general, these demonstrations have consisted of showing that genetic information, either of bacteria or viruses, could be transferred to a recipient cell by the sole means of nucleic acid.

The most complete analyses have been done in experiments with bacterial transformation.[5] Pure DNA is extracted from a bacterial suspension of one genetic type and added to bacteria having different genetic properties. Genetic characters are transferred in this way from the original cells, which donated the pure DNA, to the recipient cells. The material which produces the transformation has been purified to the point where no bacterial substances other than DNA are detectable, and the amount of protein which might be in these preparations and escape detection is less than 0.02%. That any small undetected component might have biological activity is made unlikely by the fact that, as the purification proceeds, the amount of transforming activity in the preparation is proportional to the quantity of DNA and independent of the quantity of any other components. Furthermore, it has been shown that the number of transferred cells is proportional to the amount of DNA actually absorbed by the recipient cells. Additional evidence has been provided by experiments in which enzymatic degradation is used specifically to degrade various components. Only the enzymes which degrade DNA eliminate the ability of the purified material to transform cells genetically.

The results of the experiments, which seem to be quite unambiguous, show that one can transfer genetic information from one cell to another by means of purified DNA. Thus, it can be concluded that, at least at some stage in the life cycle of the bacteria, the genetic information must be contained in DNA. However, the next question arises when the duplication of the cells is considered. Under these conditions, it is clear that the genetic material also must duplicate; the simplest hypothesis for this would seem to be that DNA of a specific type makes more of its own kind. One cannot eliminate the possibility that, in the process of duplication, the information must be passed to some other substance before it is passed back to new DNA. Although there has been a good deal of speculation along these lines,[6]

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

there are no experiments which definitively show that the information is transferred to anything but DNA.

Other experimental results which indicate that the genetic information is contained in nucleic acids have been obtained with viruses and with genetic recombination in bacteria. The first evidence that only the DNA of viruses carries genetic information came from the experiments by Hershey and Chase[7] using the bacteriophage *T*2. This is a virus which can duplicate only after it has attached to a bacterial cell. The virus particle consists of a head and a tail. The tail is composed of proteins exclusively and the head of a protein membrane surrounding DNA. When the virus attaches to a bacterial cell, it uses the tip of its tail, and after attaching, the DNA which was contained in the head of the virus is transferred into the bacterial cell, leaving the head membrane and most of the tail on the outside of the bacterium. This residual protein which remains on the outside can be sheared off by stirring the infected cell in a Waring Blendor and the bacterium, with the DNA of the virus inside it, will develop then and produce a large number of virus particles identical with the one used for infection. Thus, most of the viral protein can be eliminated, indicating that it plays no role in the formation of the new virus particle.

The most obvious interpretation of these experiments is that the protein acts as a microsyringe injecting the DNA into the host cell, and that this DNA carries the genetic information necessary for the formation of the new virus particles. On the other hand, there is a measurable amount of protein which cannot be eliminated and which does enter into the cell with the injected nucleic acid. The fact that the atoms of the associated protein do not appear in the progeny virus, while more than 50% of the atoms of the injected DNA do, also suggests here that it is solely the DNA which is the genetic material. It must be kept in mind, however, that the associated protein may play an essential role in initiating the process of growth in the invaded cell.

In the case of tobacco mosaic virus (TMV), similar results have been obtained, but a more definitive analysis has been possible. In this case, one can completely separate the protein and the nucleic acid of the virus by chemical means. It is found, here, that the nucleic acid is infective and that the protein is not.[8] If one reconstitutes the nucleic acid from one strain of TMV and the protein from another into a single particle, then the reconstituted particles have the genetic properties of the strain from which the nucleic acid was taken but not of the strain from which the protein was taken.[9] It is important to note, however, that, in the case of TMV, the nucleic acid involved is ribonucleic acid (RNA). The same kinds of analytical and enzymatic demonstration of the purity of the RNA have been carried out as were discussed in the foregoing for the transforming factor. Here, the results indicate that the genetic information is carried in the RNA exclusively. Again, as in the case of DNA, one can argue about the lower limits of detectibility of some odd component, but this is also a case where one has to push the evidence rather far to escape the conclusion that it is nucleic acid which contains the genetic information.

Genetic recombination takes place between fertile strains of bacteria by direct contact of the donor and recipient cells.[10] It has been shown that a cytoplasmic bridge is formed between these two cells and that DNA passes from the donor to the recipient. It has been demonstrated that the amount of DNA transferred is proportional to the amount of the genetic material as measured along the genetic map. These experiments have not yet been done in such a way as to eliminate the possibility of small amounts of other components also being transferred, but the apparent proportionality between the amount of DNA and the length of the genetic map again suggests that the genetic information is contained in DNA.[10,11]

Since the structure of RNA has not been worked out yet, it is difficult to say very much about the significance of those cases where RNA carries genetic information. In the case of the genetic information contained by the DNA, however, it is clear that the information must reside in the order of the four bases along the DNA chain. This is evident from the fact that, as far as the sugars and the phosphate-ester linkages are concerned, the DNA molecule is extremely monotonous with every part of it identical to every other part. One can look, therefore, at the three essential properties of any genetic material in terms of the base sequence of the DNA molecule containing the genetic information. The first activity that a genetic entity must be able to perform is reproduction and, in this regard, the suggestion made by Watson and Crick[12] as to the possible mechanism for the duplication of DNA is extremely hopeful. At present, there are no experiments to contradict this hypothesis and there are a large number which seem to be consistent with it. In addition to duplicating its own kind, genetic material must be such that recombinants can be formed. Although a great deal already is known about the way in which recombinants are formed, it will probably be necessary to know much more about the physicochemical properties of DNA in solution before a self-consistant model can be constructed with any confidence. Specifically, one must find out whether the DNA in cells has all of its possible hydrogen bonds intact or whether some regions of the molecule have two unpaired chains available for pairing with homologous parts of other DNA molecules. If this is so, then it is possible to construct a model for the recombinational event consistent with all of the available genetic data. Finally, the genetic material must be able to exercise its control; that is, it must be capable of controlling the formation of proteins and perhaps other

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

cell components with specific composition and configuration. This problem is taken up in a later article (p. 249).

BIBLIOGRAPHY

[1] C. Levinthal, "Bacteriophage genetics," in *The Viruses*, F. M. Burnet and W. M. Stanley, editors [Academic Press, Inc., New York (to be published)], Vol. II, Chap. 8.

[2] M. Chase and A. H. Doermann, "High negative interference over short segments of the genetic structure of bacteriophage *T4*," Genetics (to be published).

[3] S. Benzer in *The Chemical Basis of Heredity*, W. D. McElroy and B. Glass, editors (The Johns Hopkins Press, Baltimore, Maryland, 1957), p. 70.

[4] E. W. Sinnott and L. C. Dunn, *Principles of Genetics* (McGraw-Hill Book Company, Inc., New York, 1958), fifth edition.

[5] R. D. Hotchkiss in *Phosphorus Metabolism*, W. D. McElroy and B. Glass, editors (The Johns Hopkins Press, Baltimore, Maryland, 1952), p. 426.

[6] G. S. Stent in *Advances in Virus Research* (Academic Press, Inc., New York, 1958), Vol. V, p. 95.

[7] A. D. Hershey and M. Chase, J. Gen. Physiol. **36**, 39 (1952).

[8] A. Gierer and G. Schramm, Nature **177**, 702 (1956).

[9] H. Fraenkel-Conrat, B. A. Singer, and R. C. Williams in *The Chemical Basis of Heredity*, W. D. McElroy and B. Glass, editors (The Johns Hopkins Press, Baltimore, Maryland, 1957), p. 501.

[10] F. Jacob and E. L. Wollman in *Biological Replication of Macromolecules*, Symp. Soc. Exptl. Biol. **12**, 75 (1958).

[11] A. Garen and P. D. Skaar, Biochim. et Biophys. Acta **27**, 457 (1958).

[12] J. D. Watson and F. H. C. Crick, Cold Spring Harbor Symposia Quant. Biol. **18**, 123 (1953).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

28
Replication of Nucleic Acids

ROBLEY C. WILLIAMS
Virus Laboratory, University of California, Berkeley 4, California

SOME of the evidence for the belief that genetic information resides in the structure of nucleic acid has been presented by Levinthal (p. 227). A key question in molecular biology is the mechanism by which the replication of such genetic information takes place. It is certainly agreed that nucleic-acid replication takes place within living cells, and, in a few cases mentioned later, some information bearing on the mechanism of replication has come directly from observations on growing, normal cells. But it is also possible to observe phenomena related to nucleic-acid replication in cases where the nucleic acid of interest has been allowed to enter the cell from the outside. Three distinct systems of this nature are known: (1) Some bacterial cells release a product, called the "transforming principle" and now known to be pure DNA, that is capable of changing in an hereditary manner certain characteristics of other cells which have been exposed to the transforming material.[1] (2) Infection of bacterial cells by DNA-containing viruses (bacteriophages) causes the viral DNA to be replicated. (3) Purified RNA, notably from tobacco mosaic virus (TMV), can be mechanically inoculated into plant cells, thereby initiating the replication of more TMV-RNA encased within whole virus particles.[2] The greater part of this paper deals with experiments with bacteriophages, since it is with these systems that the greatest amount of quantitative and precise information is available.

GROWTH OF BACTERIOPHAGE WITHIN INFECTED CELLS

Although dozens of different bacteriophages are known, the group of the "*T*" phages that infect and lyse cells of *E. coli* has been worked with by far the most extensively. Within the *T*-group itself, there are three closely related strains, *T2*, *T4*, and *T6* (the *T*-even phages), which have furnished much of the biological and biochemical information that bears on nucleic-acid replication.

There are several important observed facts about the intracellular growth of the *T*-even phages.[3–5] In the extracellular form, the infective phage particle consists of a head and tail. The head consists of a thin, protein membrane, within which is stuffed all the phage DNA, amounting to about 300 000 nucleotides. The tail is solely protein and serves as the attachment organ, having a high degree of host-range specificity, and through which the DNA is injected into the cell upon infection. Bacteria usually can be infected with several phage particles, with the likelihood of success diminishing with the length of time elapsing since infection by the first particle. Except for a small amount of soluble protein, polypeptides, and polyamines, the only phage material transferred to the interior of the cell is the DNA.[5–7]

Shortly after infection the host cell knows that it is in for trouble. The synthesis of host RNA drops almost immediately to about 3% of the normal rate. Although the synthesis of total protein is unaffected, it is not long before most of it is directed toward making phage-specific structures. Total DNA synthesis is held up at first, and after about 10 min at 37°C the pre-existing host DNA begins to be transformed into phage DNA. Just before lysis, most of it has disappeared in this manner. In the case of the *T*-even phages, this depletion is made evident by the complete loss early in infection of the structural integrity of the cell nucleus.

In the meantime, there is activity in the synthesis of phage-specific materials. For about 15 min, no infectivity can be demonstrated in the contents of artificially disrupted cells,[8] this observation defining the so-called eclipse period. At the end of only about 6 min, some RNA, related to phage infection, begins to be synthesized. Since it does not appear in the mature phage, it is not a precursor material, but it does differ in its base composition from normal host RNA.[9] At about 8 min, phage-precursor DNA begins to be synthesized (and also to be converted from host DNA) so rapidly that 50 phage units of it have accumulated at the end of the eclipse period. At about the same time, various protein fractions that are phage-specific antigens begin to be made. Some of this material is not incorporated into infective phage and remains as surplus antigen. After about 15 min, it is possible to detect infective phage units in the contents of disrupted bacteria. Their number increases rapidly at first and then more slowly as the average lysis time is approached.[10] The average burst size is something like 100 phage particles, although this number varies greatly from cell to cell. Even at the time of lysis there exist within the cell about 20 phage units of excess phage DNA, the same amount of surplus phage antigen, and about 10 units of nonhost RNA. Evidently lysis does not result from a completion of phage maturation, but represents rather a catastrophic collapse of the integrity of the cell wall (see Fig. 1 for a summary of this paragraph).

There are two distinguishable periods during the infective cycle. During the eclipse period, the components of the mature phage are being synthesized and are formed into a precursor pool. At this stage, DNA replication and interchange begin to take place. Subsequently, preformed structural units begin to be as-

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

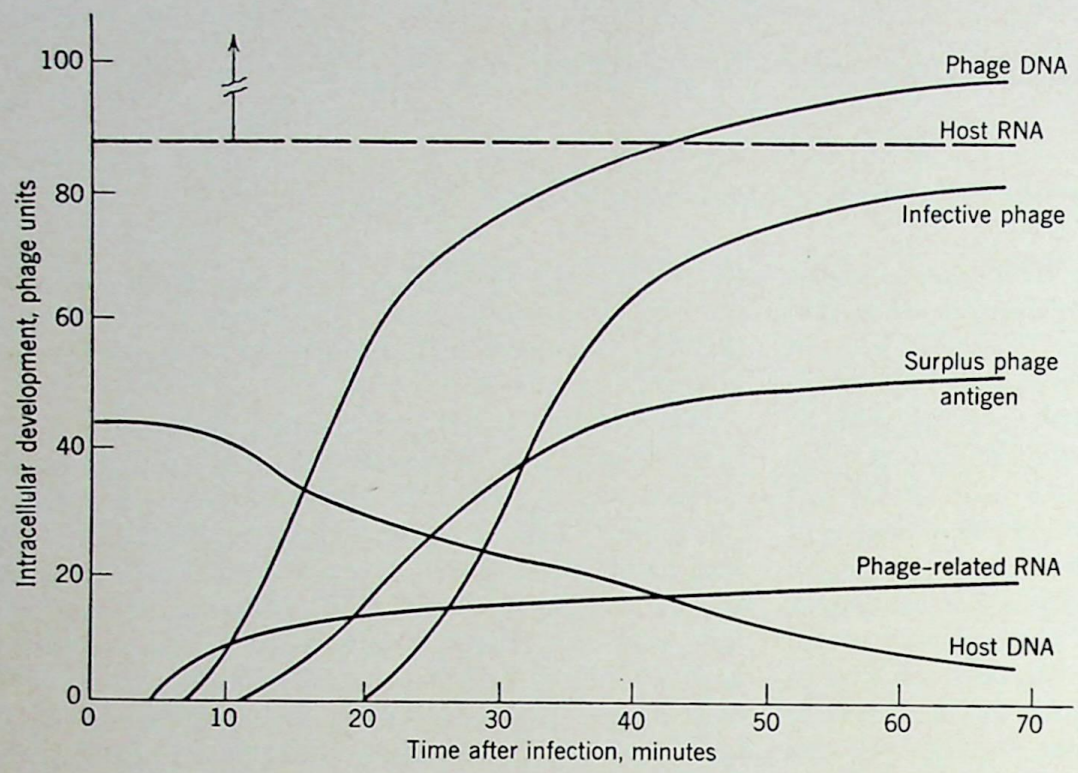


FIG. 1. A semischematic representation of the time course of the intracellular development of certain protein and nucleic-acid components following the infection of *E. coli* with *T*2 bacteriophage in a one-step growth experiment.

sembled into whole particles, although during this time some synthetic activity still persists. For our purposes, this period of maturation is of less interest than the one that precedes it, the so-called vegetative stage.

Inquiry into the origin of the DNA which ends up in the progeny phage is greatly aided by isotopic labeling, and by the fact that the DNA of the *T*-even phages can be distinguished from host DNA by the presence of the pyrimidine, 5-hydroxymethylcytosine (5-HMC).[11] It is found that *T*2-phage DNA comes mostly from material that is present in the growth medium at the time of infection; pre-existing host DNA contributes less than one-third of the total finally assembled into phages. This host DNA is evidently broken down into low molecular-weight fragments since it contains 5-HMC when repolymerized into phage DNA. As might be expected, the earliest phage to be matured are relatively rich in DNA from the host. The last phage to be matured contain some host DNA, further evidence that a large pool is formed, consisting of host DNA, newly synthesized DNA, and DNA contributed by the parent particles.

In contrast to the DNA, the phage protein is found to be synthesized solely from the growth medium. The protein that is made within 5 min of infection is found very sparsely in the progeny phage, while about half that made later is found in them. It is important to note these two kinds of protein: the nonprecursor and the precursor variety. When the formation of phage precursor protein is blocked by chloramphenicol, with the result that no progeny phage are formed, the synthesis of phage DNA goes on unabated.[12]

Although it might seem more profitable to concern ourselves with phage-precursor material only, it should not be forgotten that the kinetics of the synthesis of nonprecursor material may have a bearing on the understanding of phage DNA replication. It might be expected that the formation and fate of RNA within a cell would be unaffected by infection, since no RNA is in mature phages. But it is found that the RNA assembled after infection is unlike normal host RNA in two respects:[9] (1) its base composition is different, and (2) it suffers a rapid metabolic turnover, in contrast to the RNA of uninfected cells. This observation may be important in later considering DNA replication.

As already mentioned, two kinds of protein appear to be synthesized in infected cells: a precursor protein, and a nonprecursor one that is formed preferentially immediately after infection. Its synthesis is closely related to the formation of phage DNA, for, if protein synthesis in general is blocked by chloramphenicol during the first 5 min, no phage DNA is formed. But if blocking is delayed, phage DNA can be formed in amounts that are greater, the longer the delay. There is apparently a stage at which the assembly of protein and DNA become independent of each other, although they are severely dependent at the start. Experiments also have shown the inverse relation: that, if the assembly of phage DNA has started, its subsequent blocking by ultraviolet irradiation will not halt the synthesis of phage-precursor protein.

FATE OF PARENTAL DNA

After attachment of phage particle to bacterium, almost all of the phage DNA enters the cell. Experiments using P^{32} and C^{14} have been used to determine how much of the parentally injected DNA is found in the first and subsequent progeny generations. The experimental results with *T*-even phages are these: at each generation about one-half of the P^{32}-labeled nucleotides find their way into the progeny particles.[13]

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

(Whether or not the transfer is as inefficient as this is somewhat in doubt because of technical limitations in the recovery of all of the labeled material.) It is unlikely that the loss of the parental P^{32} atoms is due to a drastic breakdown of the phage DNA resulting in a redistribution of the P^{32} atoms among both the phage-precursor and the host DNA. Experiments of both biochemical and genetic nature strongly indicate that fairly large nucleotide sequences are transferred intact from parent to progeny. It also appears that phage DNA is not made up of a portion that is transferred with 100% efficiency, and a portion transferred not at all. If this were so, the efficiency of transfer from first-generation progeny to second-, and from second- to third-, would approach 100%, as distinct from the observed constancy of the efficiency factor. The most likely explanation is that there is a combination of apparent losses owing to technical reasons, and of random losses of fairly large phage nucleotide sequences during the process of infection, replication, and maturation.

The kinetics of DNA transfer can be followed by breaking open the infected cell at various times and examining the state of aggregation of the parental DNA.[14] During the eclipse period, such DNA is found partly in free polymers and partly in lower molecular-weight material. After the eclipse period is over, the earlier phages to mature contain more of the parental DNA, with the latest to mature containing practically none. These results are in agreement with all others which suggest that all phage-precursor substances lie around in a pool until the onset of maturation begins to withdraw them from it, apparently at random.

So far, only the content of parental DNA among the average of the progeny particles has been considered. It might be expected, however, that the distribution of the parental DNA from one individual progeny particle to another would throw more light on replication mechanisms. Two extremes are possible: (1) all of the transferred DNA winds up in one particle; (2) it is equally divided among all progeny particles. The techniques used to determine this distribution are the counting method of Levinthal and Thomas,[15,16] in which the P^{32} content of the phage particles is assayed by counting β-decay "stars" on nuclear emulsions, and the "suicide" method of Stent,[17] in which the content of P^{32} is assayed by its inactivation of those phage particles that harbor it. The former method has high precision in assessing the *size* of stars (the P^{32} content of the phages), but a low precision in assessing the *number* of phages per infected bacterium that produce detectable stars. The relative sensitivities are in the inverse order in the suicide experiments—the P^{32} content of highly labeled phage is poorly determined, while the number of phages containing at least some minimal amount of P^{32} is precisely determined.

Both methods agree in showing that the distribution of parental DNA is neither all-or-none, nor is it uniform from particle to particle. Rather, the experiments show that a very few of the first-generation progeny particles, per bacterium, individually contain as much as 20 to 25% of the parental P^{32}. The remainder of the progeny particles contain far less, with some containing no detectable amount. About half of the parental DNA goes into one first-generation progeny particle. Since Levinthal and Thomas have found that phage particles, uniformly labeled in their P^{32}, when osmotically shocked, disclose a nondispersed DNA piece containing about 40% of the original label, they have been inclined to believe that just half of this initial "big piece" stays intact and goes into the highly labeled phage particle of the first-generation progeny—i.e., it is one-half of the phage chromosome, presumably a DNA duplex. When the first-generation progeny are caused to infect, the P^{32} label continues to be transferred to the second generation, but the quantitative measurement of the results is less reliable. Levinthal finds that the star size is undiminished—i.e., that the big piece stays intact—but he cannot be certain as to the number of such stars per bacterium. Stent finds that the *number* of highly labeled phage per bacterium is the same as with the first generation, but he cannot be sure that the P^{32} content of these has remained the same. However, it appears certain that there is a highly nonuniform transfer of parental DNA; whether or not it is a single piece that could be called the phage chromosome awaits extensive and convincing genetic investigation. The idea that the big piece is the genetic template does have the merit of simplicity. It can be visualized that upon infection the big piece of the parental phage remains intact until at some stage it comes apart exactly into two halves. One half may be broken down into smaller units like the DNA units of the parental phage outside the big piece. The other half serves as a template for the assembly of other "half big pieces," two of which enter every newly matured phage. The original half big piece is preserved intact to enter one of the progeny phage (and to result in a star, if the parental phage is P^{32} labeled). About one-half of the DNA in each progeny phage will be in the form of smaller units which had as their templates either a portion of the original half big piece or the smaller DNA units from the parental phage.

DIFFERENCES AMONG PHAGE TYPES

So far, the biological and biochemical events associated with infection by only the *T*-even phages have been discussed, but it would be incorrect to conclude that they are examples of all phage types that have been studied. No comparable amount of information respecting DNA synthesis is known for the other phages for the reason that they do not contain a DNA species distinguishable from the host DNA. Only the *T*-even phage have been found to have 5-hydroxymethylcytosine (5-HMC).

Distinct differences do exist between infection with

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

TABLE I. Characteristics associated with infection by phages of different types. Most phages are found to fall into one of two classes, here called class *A* and class *B*.

Class	uv Sens.	Mult. React.	Capacity of uv cells: Normal phage	Capacity of uv cells: uv phage	Cell nucleus	Vegetative radiosens.	Recomb.	Lysog.
A *T*-even *T*5	High	High	Resistant	Resistant	Lost	Decreases with time	Frequent	Never
B λ, *P*22 *p*8, (*T*1,*T*3)	Low	Low	Sensitive	Less resistant	Intact	Constant with time	Rare	Sometimes

various phages, at any rate in a biological sense, and they are worth recounting cursorily for the insight they lend toward a proper feeling of uncertainty about phage development. In a general way, phages may be put into two categories, which can be termed *A* and *B*. The *T*-even phages are examples of *A*; phages *T*1, *T*7, λ, *p*8, and P22 are among the *B* list. These may be contrasted in several different ways[3] (Table I):

(1) The host-cell nucleus is destroyed upon infection with *A*; preserved with *B*.

(2) Upon multiple infection with homologous phages with different genetic markers, it is found that genetic exchange (recombination) is frequent among the *A* types and infrequent among *B*.

(3) Phages in the *A* category are found never to enter into a lysogenic[18] relation with the cell—i.e., to enter into a nonvirulent stage within the host cell—conferring upon it the *hereditary* property of later producing phage either spontaneously or by induction. *B* phages sometimes do become lysogenic.

(4) In experiments with ultraviolet-light inactivation of virulent phages, it is found that the phages of *A* are relatively more sensitive than those of *B*.

(5) When phages of category *A* are inactivated with ultraviolet light and then are allowed to infect cells multiply, much of their infectious activity can be restored, a phenomenon called "multiplicity reactivation." Phages of *B* are not nearly as capable of multiplicity reactivation.

(6) Phages *A* and *B* also differ with respect to the "capacity" of their host cells, defined as the ability of cells to reproduce infective phage, particularly after the cells have been damaged with agents as ultraviolet light and x-rays. For phages of class *A*, the capacity of cells is much more resistant to ultraviolet irradiation than it is for class *B*.

(7) Cells may be irradiated also while the phage is in the vegetative state, allowing measurements to be made of the radiosensitivity of the infected host to the production of infective virus. The phages of class *A* (and their host cells) have a radiosensitivity that rapidly decreases with time after infection, while those of class *B* very nearly retain their initial sensitivity.

The simplest generalized conclusion that can be drawn from the foregoing set of experimental facts is that the two classes of phages differ in the degree to which, during the vegetative stage, the genome of the infecting phage is associated with that of the host cell. It appears likely that the genome of phages *T*2 and *T*4, as examples of class *A*, replicates with considerable independence of the host, while the phages of class *B* multiply only as a consequence of genomic interaction with the host cell. A virtue of this unitary explanation of what appear to be different phenomena is that it allows the preservation of a concept that has proven fruitful in the cases of the *T*-even phages. This is the notion, given quantitative elucidation by Visconti and Delbrück,[19] that *mating* is a frequent activity in the DNA pool of the vegetative state. Mating is operationally defined as having occurred if two vegetative phages have interacted in a manner such that the probability is 0.5 that they have interchanged two distantly linked genetic loci, an occurrence detected by the production of recombinant phages. Conceptually, mating may be thought of as the pairwise physical interaction of two (or more) homologous genomic structures. The arithmetical conclusion of the Visconti-Delbrück theory is that each act of genomic *replication* is associated with a mating act, the number of such double acts being about 5 for the *T*2 phages during the vegetative phase. Strength is accorded this notion by the observation that, if lysis is inhibited and the infected bacteria opened at various times, the proportion of recombinants in the matured phages increases with time after infection—i.e., the longer the time spent in the mating pool, the greater the chances of recombination. But the experiments showing a low frequency of recombinants with such phages as λ (class *B*) have cast doubt on the correctness of the Visconti-Delbrück theory. The theory can be rescued qualitatively by assuming that mating still accompanies replication, but in the case of the class *B* phages the mating is frequently between the phage and the host genomes.

MECHANISMS OF DNA REPLICATION

The paper by Levinthal (p. 227) shows that there is good evidence that DNA can be the bearer of genetic information; that the detailed structure of one chain determines that of the other; and that presumably the structure of newly formed DNA is determined by that of the parental DNA. (It should not be inferred, however,

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

that only DNA is capable of transmitting such information, since these are examples of the apparent genetic stability of the RNA-containing viruses.) The problem to consider now is the mechanical detail of how the DNA that is in the process of being formed is so controlled in its sequence of nucleotides that it usually will be genetically identical (and presumably structurally identical) with the DNA that is acting as a template. A correlative problem is how the template DNA, either directly or through intermediaries, controls the specific assembly of amino acids into proteins, to confer upon progeny structures some of their phenotypic characteristics.

It is an attractive consequence of the two-chain structure of DNA to assume that the obvious is true: that replication is preceded or accompanied by a separation of the two chains, each of which then acts as a template for the production of its mate. This could be done by allowing the chains initially to separate at an end of the helix, forming a short-pronged Y.[13] New nucleotides then would start assembling in proper sequence along the branches of the Y, the residual helix untwisting as the polymerization of the two new chains proceeded. A difficulty with any scheme of this or similar nature is the unwinding process; how can two chains involving several hundred turns separate by untwisting, both in consideration of the energy required and the hazards of hopeless tangling? To avoid this dilemma, several schemes involving rather *ad hoc* notions of mechanical processes have been proposed, such as backbone breakages occurring at every turn of the helix during replication, but all of the schemes seem to introduce more difficulties than they remove. Actually, the energy requirements of unwinding the helix are not as great as might be imagined, if the helix is restricted to turning on its own axis. If it always turns on that axis, there is also a minimal chance of tangling. Levinthal and Crane[20] showed that the energy required per turn of the helix is very small as compared with the energy of formation of the phosphate-diester linkages per turn. It seems inevitable that the transfer of information from parental DNA to new DNA and to new, specific structures of other kinds, such as RNA and protein, involves unwinding at one or more stages.

As there seems to be no likelihood in the near future of directly determining the mechanics of replication, it is profitable to examine phenomena that indirectly might indicate what kind of mechanism is the most probable. Here some of the phage experiments are relevant. Unfortunately, the transfer experiments shed very little light on the problem; they indicate that there is some conservation of large pieces of P^{32}-containing material and some dispersal of smaller pieces among several progeny particles. A suggestion by Stent,[3] which appears attractive, formulates a sequence of events during the vegetative stage whose interpretation can point toward a mechanism of replication not mentioned above. These events are characterized chiefly by three facts: (1) phage-directed RNA synthesis begins immediately after replication; (2) the blockage of protein synthesis stops phage-DNA synthesis if it occurs soon enough, otherwise DNA synthesis proceeds in any event; (3) with *T*-even phage infection, the ultraviolet light and P^{32} sensitivity decreases with time during the vegetative phase. These facts can fit a replication scheme wherein the parental DNA structure is conserved initially, at any rate, and serves as template in the *grooves* of which is assembled a single chain of *ribonucleoprotein*. This chain subsequently unwinds off the template, leaving the latter to serve in this capacity again. The RNA-protein chain wanders off in search of a mate, which may be a chain copied from another phage-parental DNA helix or from a DNA helix of the host itself (class *A vs* class *B*). Mating consists in forming a duplex helix of two reasonably identical or homologous RNA-protein chains. The new DNA is then assembled by a process the inverse of the formation of the RNA-protein chains, wherein the two chains of the DNA are helically assembled around one of the chains of the RNA-protein duplex, the other chain disengaging during the DNA assembly. The final, RNA-free DNA is then disengaged from the template RNA-protein chain by unwinding. The RNA-protein intermediaries may be used also to serve as templates for the assembly of phage-specific protein. The handing on of the guidance of DNA synthesis to a ribonucleoprotein template would account for the production of RNA shortly after infection, and for the blocking phenomena respecting protein and new DNA, since DNA synthesis could readily proceed after a few RNA-protein templates became assembled. Further, since the secondary template is now made of all-new RNA and *protein*, one might expect it to be relatively resistant to damage by ultraviolet light and to P^{32} decay. The scheme has the further merit of predicting that the effects of slightly damaged parental (or host) DNA, such as single-chain breakage by ultraviolet light or x-rays, could be effectively overcome since all that is required is that the DNA helix hold together as a duplex. It is hard to see how single-chain breaks, spotted here and there along both chains of the helix, could lead to anything other than abortion if the original Watson-Crick replicating scheme is correct. The Stent hypothesis also provides for mating at each replication, although other schemes arrange for this too.

It does not seem that the observed facts of genetic recombination can point unequivocally to either a replicating scheme that involves DNA assembly directly upon existing DNA templates or to one involving an intermediary helical structure. It is grossly inferred by light-microscopic observations of *cellular* organisms that recombination is due to the existence of breaks and subsequent crossing-over in the dividing chromosomes. With phage, however, the genetic and biological evi-

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

dence is contrary to the expectation that crossing-over among genetic structures exists. The alternate possibility is that recombinants are produced by a mechanism of "copy choice,"[21] in which a daughter genetic structure switches from one template to a closely-adjoining but slightly differing one during its stepwise assembly. Such switches may be multiple, incorporating several genetic loci from the two mated structures serving as templates. But a hypothesis of copy-choice used in interpreting genetic recombination does not result in any unequivocal answers with respect to the *mechanism* of replication, except to enhance the notion that replication is associated with mating.

Replication schemes can be categorically distinguished (on paper) with respect to the resulting distribution of the parental DNA. If the parental DNA remains intact during all replication processes, resulting in only one parental structure and numerous other DNA's containing no parental material, a *conservative* system is said to exist. A *semiconservative* system is one in which the single DNA chains of the parental duplex retain their integrity but are physically and permanently separated during replication. The original Watson-Crick replicating scheme is a semiconservative one. A *dispersive* mechanism is one in which the atoms of the parental DNA are dispersed by some kind of fragmentation among all the daughter DNA structures. It might be hoped that experimental determinations of the kind of mechanisms so categorized would lead at least to the *elimination* of some replication schemes.

REPLICATION MECHANISMS AT THE CELLULAR LEVEL

Some autoradiographic work done by Taylor and colleagues[22] on the dividing chromosomes of root cells of the English broad bean is next to be considered. The growing cells are first fed tritium-labeled thymidine, which is known to be incorporated only in the cellular DNA, and are then observed after one or two divisions in a normal medium. Colchicine is used to preserve both generations of chromosome division within one nuclear structure. The upshot of the experiment is that the chromatid structures of these cells multiply in a way that can be interpreted as semiconservative at the level of the DNA duplex. That is, if each chromatid is thought of as consisting of numerous DNA helices (arranged in presumably some regular fashion), the helices associated with any one chromatid separate into two strands with each strand preserving its identity. At the half-chromatid level, the mechanism is conservative: the DNA chains that are associated together in any one half-chromatid stay together indefinitely.

In experiments with growing *E. coli*, Meselson and Stahl[23] have been able to investigate the type of replicating mechanism at the level of the DNA polymers. *E. coli* that have been grown on N^{15} are subsequently allowed to grow for a few divisions on N^{14}, and their DNA is examined by density equilibrium centrifugation, using a caesium-chloride gradient. A gradient can be established with a sufficiently gentle density change to distinguish structures such as DNA by their relative content of N^{15} and N^{14}. DNA containing solely N^{15} will form a band in the gradient cell that is distinctly more centrifugal than will a DNA containing only N^{14}. It is found that at first, of course, all of the *E. coli* DNA is N^{15}-containing. As growth goes on, a band midway between N^{15} and N^{14} appears, followed later by a band at N^{14}. At no time does material appear at the interband region, thus establishing that dispersive mechanisms do not operate in this system. The N^{14}:N^{15} band shows that the initial N^{15}-containing DNA helices are diluted just one-half by N^{14} during the growth cycle. The most obvious interpretation of the results is that DNA replication is semiconservative at the level of the DNA duplex. This is an elegant technique indeed and it might be hoped that it would give straight answers to the mechanism of replication of *phage* DNA. Although only a preliminary report of such investigation is available at present, it appears that the answer will not come easily. When phage previously grown on N^{14} are allowed to infect N^{15}-containing cells, in an N^{15} medium, the progeny phages can be purified and the DNA examined for N^{14}:N^{15} content. What is then seen in the gradient centrifuge is a single N^{15} band. If the cells are disrupted prematurely to expose the DNA vegetative pool, a very small amount of what appears to be an N^{14}:N^{15} hybrid DNA is found. The experiments are very likely inconclusive. If one takes the results with the purified progeny phage as they come, one would conclude that there is conservative replication. But it is easy to calculate that any N^{14}:N^{15} hybrid band which might exist would be just on the limit of detectability. So the experiments prove nothing, so far.

CRITERIA FOR ACCEPTABLE REPLICATION SCHEMES

Let us return again to the problem raised by the existence of the big piece of phage DNA, along with the implied existence of small pieces. All that is known about these structures is that a P^{32}-containing big piece is held intact during osmotic shock, and that very likely half of this big piece is passed on to individual particles of progeny phage. But it is not known how the halving of the big piece comes about during replication. It could conceivably (although improbably) be two separable half-pieces, and one of these could remain conservatively integrated as a DNA duplex during replication. And what is the role of the small pieces? Can it be that the big piece and the small pieces have different roles? Until the significance of the apparent bipartite character of phage DNA is known, it is unlikely that any replication scheme will win complete adherence.

It is tantalizing to find that something as apparently straightforward as the problem of synthesizing more DNA on a template of pre-existing DNA should present

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

so many puzzles. Let us recapitulate some of the experimental conclusions which any scheme of DNA replication should either explain or, at any rate, be consistent with. "Experimental conclusions" include presently available facts, quite reasonable inferences, and conclusions from experiments yet to be conclusively performed.

(1) Replication schemes must agree with any observations on the densities of N^{14} and N^{15} DNA, since these may well indicate whether or not replication is conservative, semiconservative, or dispersive. A conclusion based on experiments with a certain type of replicating system, such as bacterial cells, is not necessarily subject to generalization to all systems, such as phages.

(2) Provision must be made for information transfer from DNA to the synthesis of non-DNA structures, such as specific proteins.

(3) With respect to phage replication there are these specific points:

(a) The numerical equivalence of rounds of mating with rounds of replication seems to be established, especially with the *T*-even phages. It is an attractive corollary that mating and replication are not independent events, and a replicating scheme might be expected to provide for this dependence.

(b) Genetic recombination is experimentally inferred to take place by some sort of copy-choice mechanism, rather than by the occurrence of breaks and crossing-over.[21] Copy-choice recombination is reasonably believed to take place during a replicating act that occurs under the influence of mated structures. Provision must be made to have homologous templates arranged in one-to-one structural opposition, thus implying the existence of highly specific medium-range forces of interaction.

(c) If copy-choice recombination takes place in the presence of mated structures, the latter must be so mutually arranged that the newly assembled DNA can become disengaged from both mated structures without tangling and without a high degree of fragmentation of the templates, *if* these are the parental DNA duplexes.

(d) The phenomena of "marker rescue" by multiplicity reactivation[24] and cross-reactivation should be accounted for. If it turns out that parental DNA may be rescued (in its ability to be replicated) even after scission of one of the two chains, this phenomenon is to be accounted for by any replicating scheme that envisions the unwinding of parental DNA.

(e) The observed heterogeneous, but nondispersive, distribution of parental DNA in the progeny must be accounted for.

(f) Biochemical events during the vegetative phase are to be made consistent with the replication scheme. In particular, the observation is important that there is an early synthesis of phage-specific RNA and that only early blocking of protein synthesis will prevent DNA synthesis.

No present scheme of replication of phage DNA satisfies all of these requirements, and only for phages is enough known to allow schemes to be anything other than exercises in model building. Any scheme involving the disengaging of DNA chains by breaking and rejoining them is unlikely on the grounds that it implies a large degree of dispersion of the parental DNA, and it meets with difficulty in mating and recombination. The obvious semiconservative scheme based on the Watson-Crick model satisfies many of the criteria but has difficulty in handling single-chain scissions, and, in its simplified form, does not account for synthesis of phage-specific proteins. It also seems not to be sensitive to the effect of early blockage of protein synthesis. Copy-choice replication in mated DNA structures also gives difficulty, for the reasons that there seems to be no reason why complete DNA duplexes should mate and that the stereo problems of disengagement are large. The scheme of invoking a ribonucleoprotein intermediary template is consistent with many of the experimental conclusions mentioned in the foregoing. It suffers from the uncertainty as to whether it is stereochemically possible in atomic detail, nor does it make any specific predictions as to the fate of the integrated parental DNA duplex. If this remains intact throughout phage replication and maturation, a conservative type of replication is demanded. But if experiments show that phage replication is precisely semiconservative, only *ad hoc* disposition of the parental DNA structure would save this model at the moment.

REPLICATION IN RNA VIRUSES

We have concentrated so far on the problems of replication of DNA that has been introduced into cells, the best examples of which are the DNA of the transforming factor and of bacteriophage. But there exist many cases in which the genetic message must have been introduced through the medium of viruses that contain no demonstrable DNA. Actually, most known viruses, both plant and animal, contain no DNA. Although none of these RNA viruses can be observed with the quantitative precision inherent in phage work, it is known that the general facts of replication are similar. They possess genetic continuity, are found occasionally to mutate, and in one instance it appears that the character of the virus genome is controlled by the type of host in which the virus is multiplied. Their general structure is similar to that of phage, with a protein exterior and some kind of core containing the RNA.

Interest in the replicating potentialities of RNA has been enhanced by the discovery that a highly purified RNA obtained from tobacco mosaic virus (TMV) is infectious.[2,25] If a solution of this material is rubbed upon tobacco leaves in the same way as TMV is applied for assay purposes, local lesions are developed, or systemic infection is initiated. At present, pure RNA can be prepared whose infectivity per unit weight is about

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

one percent of the same weight of RNA as it exists in native virus[26] particles. The infectivity can be greatly increased by reconstituting the virus—i.e., combining the RNA with protein fragments of the virus in a way such as to re-form virus particles indistinguishable from native ones. The infectivity may then rise to a figure as high as 60%. Although reconstitution has not been achieved for any other virus, it has so far been found that the RNA from another plant virus, turnip yellow mosaic,[27] is infectious, as is the RNA from polio virus.[28]

The findings just described allow RNA to be studied in a way that has some aspects of both phage and transformation experimentation. Like the transforming factor, RNA can be biologically assayed after removal from its natural host, and after purification and laboratory manipulation. Like phage DNA, its activity is demonstrated by the production of copies of the virus originally containing the nucleic acid. It is frustrating, however, that plant viruses cannot be studied with a quantitative precision even approaching that enjoyed by the phage experiments.

The work with infectious RNA has largely taken the lines of finding out what happens when artificial "hybrid" viruses are reconstituted, and of deducing the physicochemical characteristics of the infectious unit. As might be expected, when RNA from a virus of one strain is reconstituted with protein from another strain, the progeny particles are copies of the virus that furnished the nucleic acid. The protein evidently serves primarily as a protecting coat to preserve the integrity of the RNA during inoculation. It is also possible to make what might be called artificial recombinants: reconstituted TMV particles that are believed to contain nucleic acid from two parent strains.[2] When such virus particles are assayed on local-lesion hosts, most of the lesions found resemble either those normally caused by one or the other nucleic acids, but some have a mixed character. When the latter type of lesion is subinoculated by using it as inoculum for systemic infection, the symptoms developed and the character of the progeny virus from the systemic infection are wholly those normally associated with either one strain of RNA or the other. If the reconstituted particles have two kinds of RNA packaged within them, the process of infection evidently allows each kind of RNA to be independently replicated, a not surprising result.

The experiments designed to find the minimum weight of RNA that is infectious are obviously important in their bearing on the question as to whether or not there is redundancy in the amount of information-carrying material in a viral genome. Great interest would attach to the finding, for example, that a structural unit of considerably less mass than all of the RNA of the virus is sufficient to act as the replicating template. Is there a TMV-RNA big piece and are there little pieces that are infectious? The answers to these questions, if they are positive, should come more directly with the RNA from TMV than they are likely to come with the phages. So far, the results are equivocal; RNA is not easy to fractionate into differing sizes nor is it easy to characterize physically. At present, it appears that RNA pieces smaller than the RNA content of one TMV particle are infectious.

It seems that, at present, the importance of the finding that purified RNA is virogenic is a conceptual one. Even though RNA viruses have been recognized for a long time, there has been no compelling reason to believe that their infectious proclivity is localized in their RNA, in analogy with the phage-injected DNA. Now it appears that the biochemical aspects of the earliest stage of infection are similar—that in both cases reasonably pure nucleic acid can act in the initiation of infection and the messenger of genetic information. Any master plan of replication, therefore, must show how *RNA* can direct the formation of its copies, as well as how *DNA* can do this. It certainly will not be surprising to find that some kind of protein synthesis, acting in specific and intimate contact with the nucleic acids, is a necessary correlate in the whole replication process. Nor will it be surprising to find that RNA is involved in DNA replication, and that DNA is involved in RNA replication.

Where have these extended but inconclusive remarks on the replication of nucleic acids as evidenced by the viruses led us? My own feeling is that it is currently premature and naive to expect that the geometrical nicety implied in the two-strand helix of DNA will allow us to get out our poppet beads and come up with a jigsaw puzzle answer to the way such structures are assembled within cells. It is not known even that DNA and RNA *are* replicated, in the sense, for example, that the newly formed nucleic acids of genetically identical progeny viruses are identical on atomic dimensions with the parental nucleic acids (among which, of course, there may be diversity from parent to parent). Model building is fun, and its results are provocative and suggestive of what experiments to do next, but we are a long way from being able to identify the model with reality. We do not know how nucleic acids replicate, even in terms of quite nebulous mechanisms. The danger of premature model building is that it tends to direct our experimentation along subjectively controlled lines, perhaps to blind us to the key experiments that do not currently fit our preconceptions.

BIBLIOGRAPHY

[1] H. Ephrussi-Taylor in *The Chemical Basis of Heredity*, W. D. McElroy and B. Glass, editors (The Johns Hopkins Press, Baltimore, Maryland, 1957), p. 299.

[2] H. Fraenkel-Conrat, B. A. Singer, and R. C. Williams in *The Chemical Basis of Heredity*, W. D. McElroy and B. Glass, editors (The Johns Hopkins Press, Baltimore, Maryland, 1957), p. 501.

[3] G. S. Stent, Advances in Virus Research **5**, 95 (1958).

[4] G. S. Stent in *The Viruses: Biochemical, Biological and Biophysical Properties*, F. M. Burnet and W. M. Stanley, editors [Academic Press, Inc., New York, (to be published)], Vol. I, Chap. 3.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

[5] S. E. Luria in *The Viruses: Biochemical, Biological and Biophysical Properties*, F. M. Burnet and W. M. Stanley, editors [Academic Press, Inc., New York (to be published)]. Vol. I, Chap. 8.

[6] A. D. Hershey and M. Chase, J. Am. Physiol. **36**, 39 (1952).

[7] A. D. Hershey, Advances in Virus Research **4**, 25 (1957).

[8] A. H. Doermann, J. Gen. Physiol. **35**, 645 (1952).

[9] E. Volkin and L. Astrachan, Virology **2**, 149 (1956).

[10] C. Levinthal and H. Fisher, Biochim. et Biophys. Acta **9**, 419 (1952).

[11] G. R. Wyatt and S. S. Cohen, Nature **170**, 1072 (1952).

[12] A. D. Hershey and N. E. Melechen, Virology **3**, 207 (1957).

[13] M. Delbrück and G. S. Stent in *The Chemical Basis of Heredity*, W. D. McElroy and B. Glass, editors (The Johns Hopkins Press, Baltimore, Maryland, 1957), p. 699.

[14] G. S. Stent and N. K. Jerne, Proc. Natl. Acad. Sci. U.S. **41**, 704 (1955).

[15] C. Levinthal and C. A. Thomas, Biochim et Biophys. Acta. **23**, 453 (1957).

[16] C. Levinthal, Proc. Natl. Acad. Sci. U.S. **42**, 394 (1956).

[17] G. S. Stent, J. Gen. Physiol. **38**, 853 (1955).

[18] F. Jacob, *Les Bactéries lysogènes et la notion de provirus* (Masson et Cie., Paris, 1954).

[19] N. Visconti and M. Delbrück, Genetics **38**, 5 (1953).

[20] C. Levinthal and H. R. Crane, Proc. Natl. Acad. Sci. U. S. **42**, 436 (1956).

[21] J. Lederberg, J. Cellular Comp. Physiol. **45**, Suppl. 2, 75 (1955).

[22] J. H. Taylor, P. S. Woods, and W. L. Hughes, Proc. Natl. Acad. Sci. U. S. **43**, 122 (1957).

[23] M. Meselson and F. W. Stahl, Proc. Natl. Acad. Sci. U. S. **44**, 671 (1958).

[24] S. E. Luria, Proc. Natl. Acad. Sci. U. S. **33**, 253 (1947).

[25] G. Schramm and A. Gierer, *Cellular Biology, Nucleic Acids, and Viruses* (Special Publications of the New York Academy of Sciences, 1957), p. 229.

[26] H. Fraenkel-Conrat and R. C. Williams, Proc. Natl. Acad. Sci. U. S. **41**, 690 (1955).

[27] R. L. Steere (unpublished observations).

[28] J. S. Colter, H. H. Bird, A. W. Mayer, and R. A. Brown, Virology **4**, 522 (1957).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

29
Genetic Fine-Structure Analysis

E. S. LENNOX

Department of Chemistry and Chemical Engineering, University of Illinois, Urbana, Illinois

THE question of concern here is: To what extent can properties of the genetic map be made to correspond to some physical structure in the organism? Since the bulk of the evidence points to deoxyribonucleic acid (DNA) as the molecule involved, this becomes a question of refining genetic experiments to the point where they can critically examine properties of DNA which are involved in its acting as genetic material.

The experiments discussed herein are those involving genetic measurements only—essentially, recombination frequencies in certain standard kinds of crosses (see Levinthal, p. 227). Experiments attempting to correlate genetic properties with synthetic capacities of cells are discussed in the second paper by Levinthal (p. 249).

One of the most fruitful areas of investigation regarding this problem has been that of detailed fine-structure analysis. In essence, the method has been that of pushing recombination frequencies to such a limit that sizes of genetic units may be compared with the molecular subunits of nucleic acid, the nucleotides. It is a comparison between the fine-grain structure of the genetic map and the fine-grain structure of DNA.

Much of the following discussion of fine-structure analysis is based upon experiments by Benzer on the bacteriophage $T4$.[1] Discussed also are experiments with bacteria bearing on the question posed above.

The reasons for choosing microbial systems for these kinds of studies are several. Mutants are available in large numbers and varieties, and selective techniques make it possible to screen large numbers of organisms in the search for a desired type. For example, in the $T4$ system (discussed in detail in the following), one wild-type bacteriophage particle can be detected among 10^6 to 10^7 mutant types, if the proper bacterial strain is used as an indicator. This allows examining recombination frequencies of the order 1 in 10^6 to 10^7. Another reason for the choice of microbial systems is that homogeneous preparations can be made with relative ease for chemical studies.

The work of Benzer made it necessary to define carefully the genetic units in terms of genetic measurements. The classical unit of heredity, the gene, has several operational definitions which had, in general, not been distinguishable experimentally in *Zea mays* and in *Drosophila*, the principal organisms of research in genetics. The definitions are as follows: (1) The unit of recombination—the *recon*—is defined as the smallest element interchangeable in recombination experiments. In such experiments, one uses easily-recognizable characters which enter the crosses in the genetic material of the parents and which sort among parental-type or recombination-type progeny. This element is not necessarily the same as that defined by another set of operations—isolation of mutant types (either spontaneous or induced) and mapping of their sizes. This yields another unit. (2) The unit of mutation—the *muton*—is the smallest element, alteration of which gives a type recognizably different from the standard form chosen as wild type. What is recognized as a mutant depends on the method of observation. It is, of course, conceivable that different mutant characters will be of different sizes. (3) The last unit that needs definition is the unit of function—the *cistron*. This unit is more difficult to define because what is recognized as the function of the genetic material depends upon the level at which observation is made between the primary synthetic act of the genetic material and an observable result of this act. If, for example, wing shape in *Drosophila* is the character being scored, then a large number of steps leading to the mutant form may be involved. One needs a way of deciding that two mutants are blocked in the same function, in the absence of knowing what the precise function is in terms of primary genetic product. A definition of a functional unit by genetic experiments is provided by the (*cis*-)(*trans*-) test devised originally for *Drosophila* by Lewis,[2] and used so elegantly by Benzer in his study of $T4$ bacteriophage. One asks the question: Are two given mutants with the same outward appearance affected in the same functional unit? This is answered by allowing the genetic material of both mutants simultaneous expression in the same cell. In *Drosophila*, this is accomplished by introducing one mutant into the egg and the other into the sperm. The resulting cells then have the two mutants' markers in homologous but not identical sets of genetic material. In the bacteriophage system, the test is performed by simultaneous infection of the host bacterium with two mutant particles. In each case, one wants to know whether the resulting combination (the *trans*-configuration) is wild type or mutant in its functioning. The control experiment is the similar situation in which both mutants (as a double mutant) are in one complement of genetic material, the other complement being wild type (*cis*-configuration). What are the expected results of such experiments? First the control: this must function as wild type if the other tests are to be revealing. If, in the *trans*-test, both mutants are in the same functional unit, the combination is not expected to be competent. If, on the other hand, they are mutant in different functional regions, then the combination should be competent, since each can make up for the deficiency of the other. Note that

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

the competence does not depend upon recombination to give wild type, for the genetic material would in subsequent tests still be revealed as mutant. Two mutants then are said to belong to the same functional unit—i.e., the same *cistron*—when in the *trans*-configuration they function as mutants.

Consider in some detail the system exploited by Benzer. The bacteriophage mutants used were many independent isolates from wild-type parents which show different plaque morphology (here designated *r*-type) on the host strain *E. coli B*. On other bacterial strains, these mutants have properties which are compared with wild type in Fig. 1.

The classification of mutants into groups *rI*, *rII*, and *rIII* is done on the basis of the characteristics indicated in Fig. 1. The group of mutants *rII* is especially interesting in not giving plaques on strain *K*. This allows in a cross selecting against two parental *rII* mutants and selecting for rare wild-type recombinant.

Figure 2 shows how the appearance of the genetic map depends upon the choice of the bacterial host for examining the bacteriophage mutants. The designation of *rII* on *K* as *m* comes from the fact that an infrequent

Phage strain	Bacterial host strain		
	B	*S*	*K*
wild	wild	wild	wild
rI	*r*	*r*	*r*
rII	*r*	wild	...
rIII	*r*	wild	wild

FIG. 1. Plaque morphology of *T*4 strains (isolated in *B*) plated on various hosts [from S. Benzer in *The Chemical Basis of Heredity*, W. D. McElroy and B. Glass, editors (The Johns Hopkins Press, Baltimore, Maryland, 1957), p. 70].

rII mutant gives a minute turbid plaque on this host—most of them, however, give no plaques. In the physiological effect of the mutational event, these mutants vary, then, from lethal to unnoticeable, depending on the host used for phage assay.

Turn now to the problems of measuring, in terms of recombination frequencies, the sizes of the genetic units defined in the foregoing.

To measure the unit of recombination, the appropriate bacterial cell (here, *E. coli B*) is infected with two parental *rII* mutants. One measures the frequency with which this pair of mutants recombine to give wild type (assayed on *K*). By searching for many *rII* mutants and by mapping each against the others, one sees if there is a limiting distance between positions at which two mutational events occurred.

To measure the muton length, one observes whether, with three closely linked markers, one can find aberrations in recombinant frequencies owing to the size of the mutation (Fig. 3). This size is reflected by a nonadditivity of distance 1 to 2 and 2 to 3 as compared with 1 to 3.

To determine a cistron length, many isolates of *rII*-type mutants are tested in pairs in a *trans*-

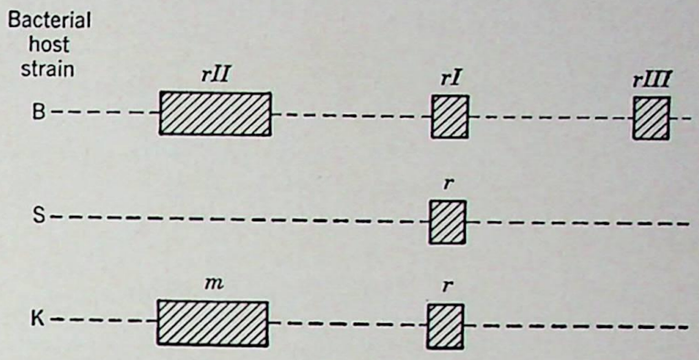


FIG. 2. Dependence of the genetic map of *T*4 upon the choice of host. Three regions of the map are shown as they probably would appear if *E. coli* strains of *B*, *S*, or *K* were used as the host [from S. Benzer in *The Chemical Basis of Heredity*, W. D. McElroy and B. Glass, editors (The Johns Hopkins Press, Baltimore, Maryland, 1957), p. 70].

configuration; e.g., *rII* mutant *m* and *rII* mutant *n* are put into the same cell, and this combination scored as wild type—i.e., yields phage, or as mutant, i.e., no phage yielded. This establishes whether the two mutants are in the same cistron. Two mutants in the same cistron are then mapped relative to each other and to other members of the cistron. The largest distance between two mutational markers within the same functional unit gives a lower bound to the size of this unit.

The amount of labor involved in mapping the numerous mutants necessary for such studies is considerably reduced by mapping each newly discovered mutant against a set of deletions that span the regions to be mapped. Mutants labeled as deletions behave as if a portion of the genetic map were missing. They do not revert to wild type at a measurable frequency. One can find for any deletion a group of reverting mutants which among themselves yield recombinant types, but none of which do so with the given deletion. The extent of the deletion covers this group of reverting mutants. The reverting mutants themselves are mapped as points on the map.

The enormous advantage of using these deletions is that they span blocks of the map, allowing any new mutant to be located quickly without an excessive number of crosses. Detailed mapping of a newly found mutant is then done only with regard to those pre-

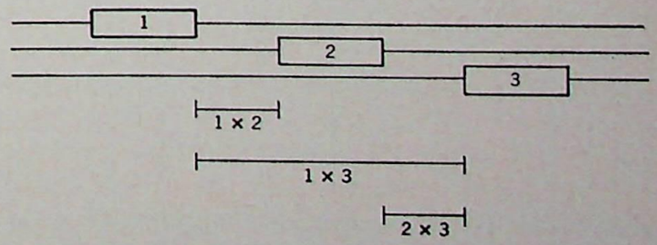
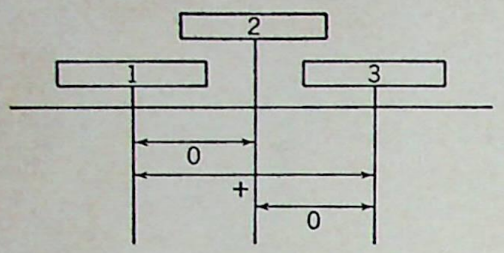


FIG. 3. Method for determining the "length" of a mutation. The discrepancy between the long distance and the sum of the two short distances measures the length of the central mutation [from S. Benzer in *The Chemical Basis of Heredity*, W. D. McElroy and B. Glass, editors (The Johns Hopkins Press, Baltimore, Maryland, 1957), p. 70].

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA



	1	2	3
1	0	0	+
2	0	0	0
3	+	0	0

(b)

FIG. 4. The method of overlapping deletions. Three mutants are shown, each differing from wild type in the deletion of a portion of the genetic material. Mutants No. 1 and No. 3 can recombine with each other to produce wild type, but neither of them can produce wild recombinants when crossed to mutant No. 2. The matrix (b) represents the results obtained by crossing three such mutants in pairs and testing for wild recombinants; the results uniquely determine the order of the mutations on the map (a) [from S. Benzer in *The Chemical Basis of Heredity*, W. D. McElroy and B. Glass, editors (The Johns Hopkins Press, Baltimore, Maryland, 1957), p. 70].

viously discovered ones that lay within the same deletion.

Mapping of the deletions with regard to each other is also done relatively simply (Fig. 4). In the figure, mutants 1 and 2, for example, will not give wild type in a mixed infection and are defined as overlapping deletions. On the other hand, mutants 1 and 3 will give wild-type recombinants and are nonoverlapping deletions.

By these methods, Benzer was able to focus his attention on a relative small subregion of the *rII* region of the *T*4 map. By mapping each new mutant against a fixed standard deletion, it was screened either as one to be ignored for the moment (outside the deletion) or as one to be mapped further—first, against smaller subdeletions, and, finally, with the mutants in each subblock defined by the smallest deletions. In this way, refined genetic analyses were performed making possible measurements of the genetic units defined earlier.

One of the first results to come out of such a program of mapping is that the *rII* region defined originally in terms of plaque-forming characteristics on the bacterial strains *B*, *K*, and *S* consists of two functional regions, the *A*-cistron and the *B*-cistron.

Restricting attention further to one cistron allows more-refined mapping. Figure 5 indicates the results of

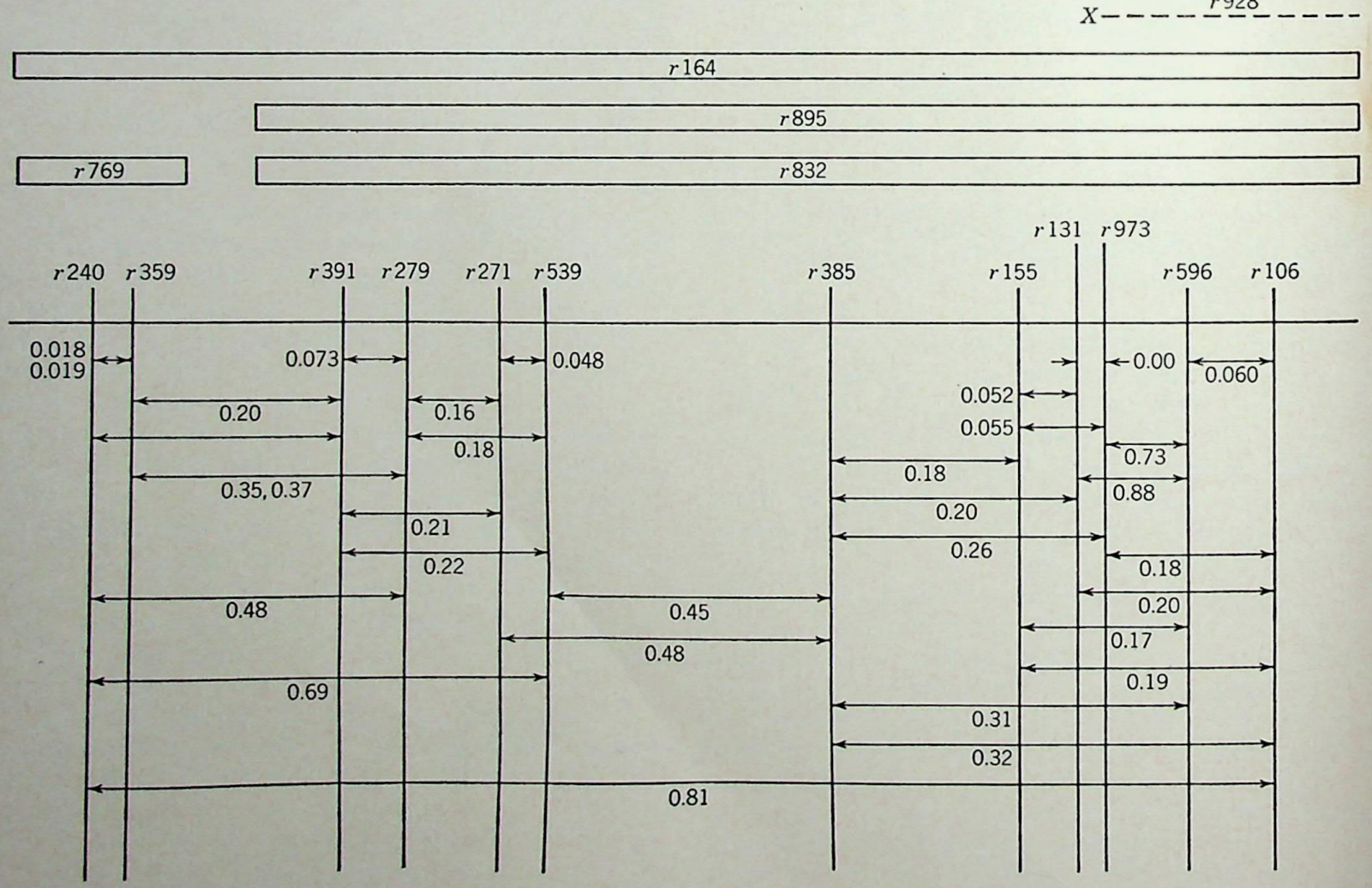


FIG. 5. Map of the mutants in the *r*164 segment. The numbers give the percentage of recombination observed in standard crosses between pairs of mutants. The arrangement on this map is that suggested by these recombination values; it has not yet been verified by three-point tests. Stable mutants are represented as bars above the axis; the span of the bar covers those mutants with which the stable mutant produces no detectable wild recombinants. The stable mutant *r*928 appears to be a double mutant having one mutation at the highly mutable *r*131 location and a second mutation at a point in the *B*-cistron. Mutants *r*131 and *r*973 are separated on the map so that the data for each can be indicated. Some of the data here given differ from (and supersede) previously published data based upon unconventional crosses which turned out to be incorrect [from S. Benzer in *The Chemical Basis of Heredity*, W. D. McElroy and B. Glass, editors (The Johns Hopkins Press, Baltimore, Maryland, 1957), p. 70].

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

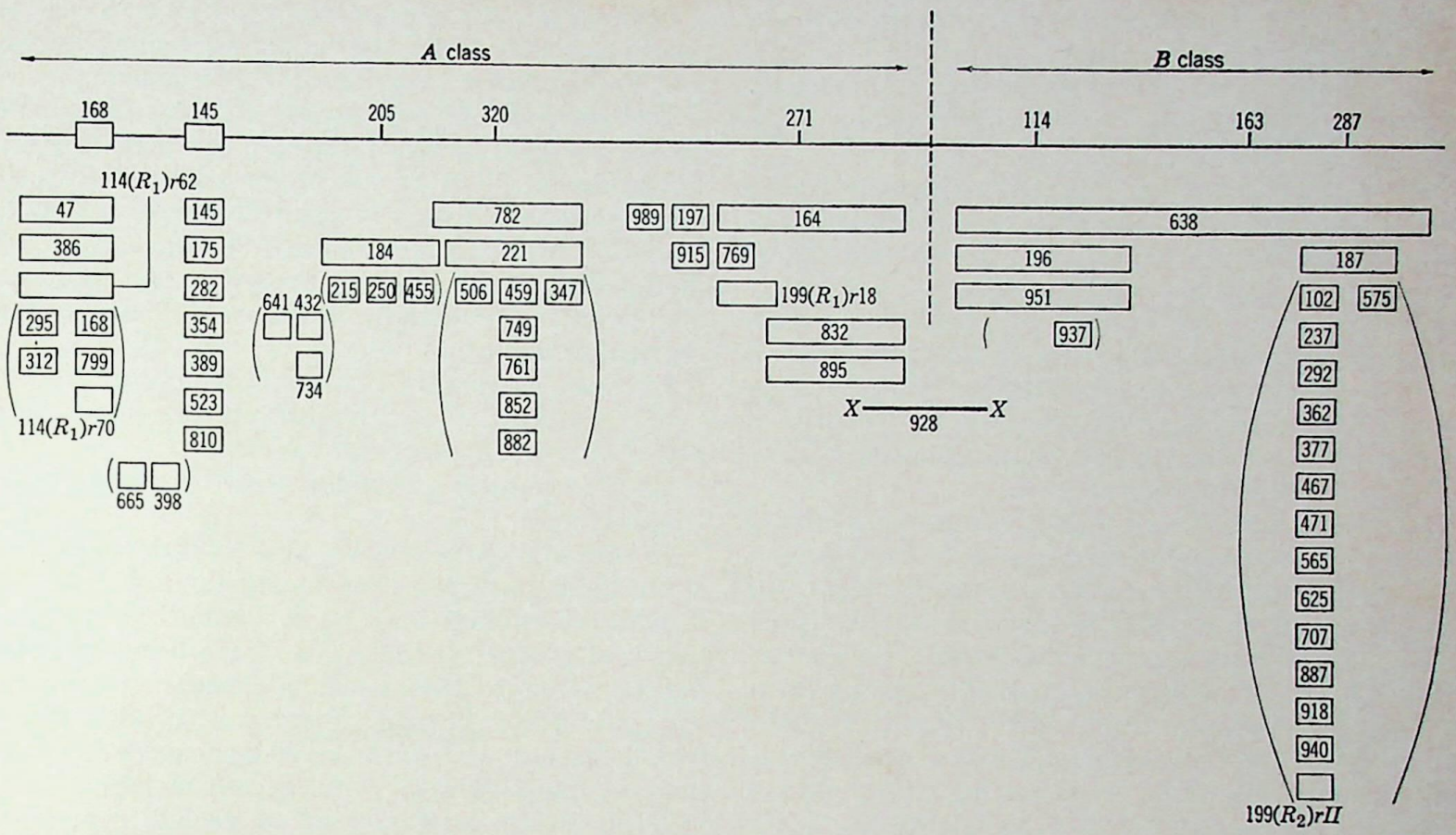


FIG. 6. Preliminary locations of stable *rII* mutants. Mutants producing no wild recombinants with each other are drawn in overlapping configuration. Pairs which produce small amounts are placed near each other. Since there remain some gaps, the order shown depends upon that established by Doermann (in personal communication with S. Benzer) for the mutants shown on the axis. The scale is somewhat distorted in order to show the overlap relationships clearly. Brackets indicate groups, the internal order of which is not established. Ten stable mutants of the *A* class and six of the *B* class were not sufficiently close to any others to permit them to be placed on the map. (*A* class equals *A*-cistron; *B* class equals *B*-cistron) [from S. Benzer in *The Chemical Basis of Heredity*, W. D. McElroy and B. Glass, editors (The Johns Hopkins Press, Baltimore, Maryland, 1957), p. 70].

such detailed mapping in a restricted region of the *A*-cistron. The number indicative of the size of the recon is about 0.02%, the closest distance between any two mutants in this region. It is seen in the following how this recombination frequency can indicate a size of the recon in molecular terms.

The recombination map as shown in Fig. 5 is far from filled, in the sense that very little of the total span has mutants crowded together as close as the above-indicated apparent limiting distance. Verification of this as a limiting distance will depend upon finding many more mutants to fill up the map.

The next step was to see if the map could be saturated using the deletions already discussed. Can the whole region of the *A*-cistron, concerned apparently with a single function, be spanned by the deletions? Also, can these deletions be ordered into a linear array that does cover region *A*? The overlapping of two deletions is indicated by the fact that each may yield recombinants with one-third (nonoverlapping) deletion, but not with each other. The ordering of a triplet of deletions is done in the standard way by comparing the frequencies of wild-type recombinants of each pair of the triplet. Figure 6 shows the result of the ordering of a large number of such deletions in the *A*- and *B*-cistrons. Two facts stand out. Firstly, each deletion, with one exception which is yet to be explained, falls clearly into either the *A*- or the *B*-cistron; i.e., control of the same function falls into a well-bounded map region. There were no *a priori* reasons why *A*- and *B*-mutants should not have been scattered among each other in the same region. Secondly, the array of mutants can be arranged in a unique sequence along a line.

Both of these facts and the one referred to in the foregoing, that there seems to be a smallest separation found between two mutants, are the heart of the interpretation which correlates the functioning of the genetic material with a molecular structure.

The lengths (in terms of recombination frequencies) of genetic units were found to be (1) the recon ~0.01%, (2) the muton ~0.05%, and (3) the cistron ~4%. These are not precise figures, for clearly the map has not been saturated sufficiently with markers to be confident of them. Probably the one in largest error is the muton.

In a rough calculation, Benzer translated these units into numbers of nucleotide pairs. The total amount of DNA in a phage particle has been well measured. Knowing roughly the total recombination frequency from genetic experiments and with the following assumptions, one can make the appropriate calculations:

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

(1) All of the DNA is genetic and corresponds to one copy of the genetic information.

(2) The map is uniform; i.e., a given recombination frequency anywhere on the map corresponds to the same amount of DNA.

With these assumptions, the ratio of total nucleotide pairs to recon size gives the order of magnitude 1. The limitation and expected variation in this calculation are discussed in the paper by Benzer.[1]

Consider next experiments with bacteria, investigating the correspondence between molecular properties of DNA and the genetic map. The question here is the same as for the bacteriophage system: Is there a physical structure in bacteria whose properties correspond to the genetic structure?

Bacterial strains *E. coli* exhibit two primary mating types which contribute asymmetrically to the formation of zygotes from which recombinants are derived.[3] One type, *Hfr*, acts as a donor of material and the other type, F^-, acts as the acceptor. In general, the characteristics of the recombinants formed by a mating event are derived principally from the F^- parent. The *Hfr* parent contributes only a part of the total genetic map.

Conjugation of these two mating types is accompanied by formation of a tubule between the bacteria.[4] The sequential character of genetic transfer is revealed by experiments in which mating bacteria are subjected to high shearing forces in a Waring Blendor to interrupt the mating process. This interruption, at different intervals after initiation of mating, allows the formation of zygotes with varying amounts of the genetic map from the *Hfr* parent. This process reveals a sequence of genetic markers and distances among them closely corresponding to those found by the more usual genetic crosses. Other experiments indicate that the effect of the Blendor is on the genetic-transfer process[3] rather than at later stages of incorporation of the genetic material.

To study how this genetic transfer is correlated with a material transfer, Fuerst, Jacob, and Wollman (referred to in reference 5) used P^{32} labeling in the *Hfr* parent. Earlier experiments[6] had shown that bacteria having high specific-activity P^{32} lose with time their ability to multiply. This loss varies with the number of P^{32} decays that have occurred in such a way as to indicate that each decay has a constant probability of rendering the cell incapable of survival. Two related experiments were done by Fuerst, Jacob, and Wollman. (1) *Hfr* bacteria were grown with high levels of P^{32}, washed free of excess P^{32}, frozen to stop metabolic activity, and stored frozen. At regular intervals, samples were thawed, mixed under standard conditions with an F^- strain, and mating was allowed to occur. They measured the probability that a given genetic length of chromosome be incorporated in the recombinants as a function of the total amount of P^{32} decay in the *Hfr* parent. (2) *Hfr* bacteria were grown with high levels of P^{32}, allowed to conjugate with the F^- parent to form zygotes which are then frozen, and stored as above. These are thawed at regular intervals, and the number of recombinants determined.

The two types of experiments differ in that, in the latter, injection occurs before P^{32} damage, whereas, in the former, P^{32} decay acts in destroying the capacity of a given map length to be transferred and to be incorporated in the recombinants. In the second type of experiment, the effect on recombination frequency is only on the probability of incorporation of a given segment of map.

In these experiments, two results were clear:

(1) The rate of inactivation of the genetic map varies with the length of map selected.

(2) The transfer probability varies according to the length of map.

In both of these results, the kinetics of decay indicate a constant probability of destruction of both the transfer probability and the incorporation per given length of genetic segment—the length being measured by the previous timed-injection Blendor experiments. This correlation of rate of inactivation with genetic length fits very well with the assignment of the phosphorus containing DNA as the genetic molecule.

There is still another kind of genetic experiment performed with bacteria that bears on the correlation of genetic proximity and functional relatedness. This is concerned with the mapping of genetic units controlling related biochemical functions.

The bacterial systems investigated have the advantage over the bacteriophage system in that it is easy to select for mutations controlling production of a specific enzyme catalyzing a single chemical step. With the bacteriophage system, it is easy to ascertain whether two mutants belong to the same functional region [by the (*cis*-)(*trans*-) test], but it has, so far, not been possible to ascribe synthesis of a specific protein as the function of such a region. With bacteria, finding the function is easy, but confirming that this function corresponds to a cistron as defined by Benzer has been more difficult.

In these experiments, a group of biochemically deficient mutants is isolated from a standard wild type, characterized with regard to the specific nature of the biochemical block, and each mutant is mapped.

Consider as a specific example the mapping of histidine-deficient mutants in *Salmonella typhimurium*.[7] The wild-type organism can grow on a synthetic medium using glucose as a source of carbon and ammonia as a source of nitrogen. From this wild-type parent, Hartman[7] isolated a large number of mutants having in common a requirement for histidine, in addition to the medium supporting growth of the wild-type organism. The synthesis of histidine consists of many known steps.[8] These mutants were characterized as to which particular steps were blocked—presumably through failure to make the necessary protein for catalysis of this step, or through making a

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

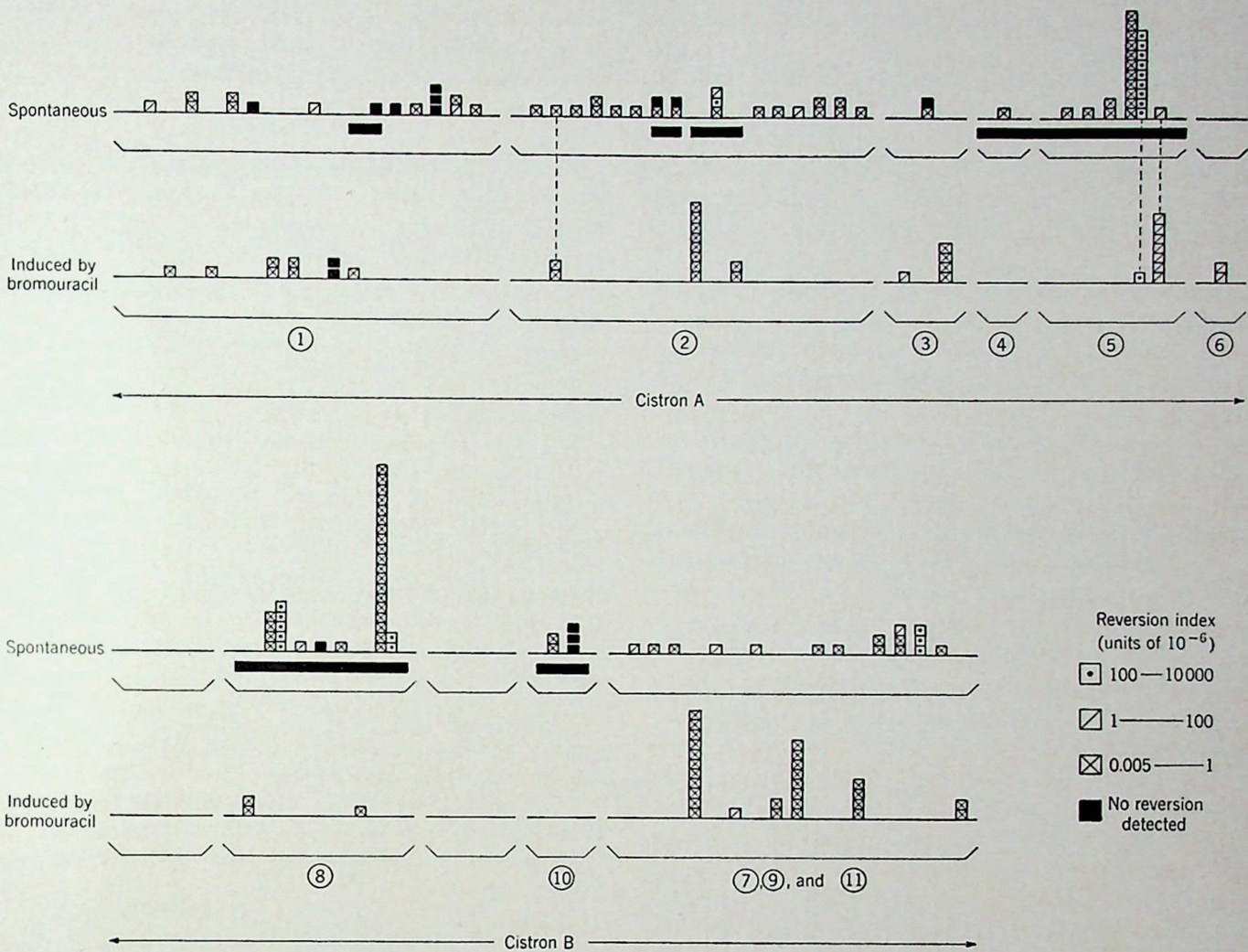


FIG. 7. Genetic maps showing the locations of spontaneous and 5-bromouracil-induced mutational alterations in the *rII* region of phage *T*4 divided into cistron *A* and *B*. Each mutant is represented by a box, placed in the proper segment of the map and shaded to indicate its reversion index. The relative order (but not the lengths) of segments 1, 2, 3, 4, 5, 6, 8, and 10 have been established. Segments 7, 9, and 11 are lumped together and considered as one segment. The order and position of different mutations within one segment have not been determined [from S. Benzer and E. Freese, Proc. Natl. Acad. Sci. U. S. 44, 112 (1958)].

protein with a defect, rendering it incapable of catalyzing the reaction step.

Comparing two histidine-requiring mutants, biochemically, reveals that they are blocked in the same or different biochemical step. Mapping of these mutants revealed two facts:

(1) All of the biochemical blocks in histidine synthesis are close to each other on the map.

(2) Those markers involved in blocking at the same step are in clusters; the markers from different steps seem to belong to separate but nearby clusters.

It is, of course, natural to speculate that such clusters are the cistrons defined by Benzer, though his defining (*cis*-)(*trans*-) test is lacking for this system.

There are many other cases in the mapping of biochemical blocks in bacteria where similar evidence has accumulated.[9–11] In all of these, the clear definition of a cistron is lacking. The experiments of Jacob and Wollman on the formation of partial zygotes by interrupted mating make such tests possible.

To summarize, experiments mapping functional blocks in bacteria have strengthened the idea that a unit function is under the control of a small region of the genetic material.

The last experiments discussed here return attention to *T*4 bacteriophage where mapping had been carried to a degree of fineness that the recombination and mutation units could be defined in terms of relatively few nucleotides.[1] This opened a new avenue to the study of mutation and to bringing into correspondence the properties of DNA with those of the genetic map.

Benzer had made a profile of the mutation frequency in various portions of the *A*- and *B*-cistrons. The spontaneous mutation frequency is not uniform throughout the map, but does have local "hot spots" of high mutation frequency. Benzer and Freese[12] then investigated whether a mutagenic agent which could be expected to interfere with the incorporation of a particular nucleotide into DNA would alter the profile of mutation. For this purpose, they used 5-bromouracil, which had been shown to be incorporated into the DNA

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

of phage replacing thymine[13,14] and to be mutagenic in bacteriophage.[15]

Since the spontaneous mutational events probably arise from alterations at more than one kind of base site, and since the mutagenic agent 5-bromouracil would be expected to interfere principally by replacing thymine, then the profile of induced mutation would be expected to be very different from the spontaneous one.

A comparison is shown in Fig. 7, which plots the mutational frequency in various segments of the *A*- and *B*-cistrons. The profile of spontaneous mutants is markedly different from the 5-bromouracil induced ones.

Such studies, extended to other mutagenic agents, may form a basis for studies of mutagenesis on a molecular level.

The experiments discussed here—on bacteriophage and on bacteria—have yielded the following observations:

(1) The genetic map is linear down to its smallest dimensions.

(2) Portions of the map controlling the same function are compact and not intermixed with portions controlling other functions.

(3) The units of recombination and mutation have sizes of the order magnitude of a few nucleotide pairs; the size of the unit of function is the order of magnitude of 10^2 to 10^3 pairs.

(4) There is a correspondence between map distance and phosphorus content of the genetic material.

This summary is consistent with the identity of DNA and the genetic material. The real problems start here—to specify at a molecular level the detailed structure of DNA in its acting as genetic memory and in its role of making genetic products.

BIBLIOGRAPHY

[1] S. Benzer in *The Chemical Basis of Heredity*, W. D. McElroy and B. Glass, editors (The Johns Hopkins Press, Baltimore, Maryland, 1957), p. 70.

[2] E. B. Lewis, Cold Spring Harbor Symposia Quant. Biol. **16**, 159 (1951).

[3] E. L. Wollman, F. Jacob, and W. Hayes, Cold Spring Harbor Symposia Quant. Biol. **21**, 141 (1956).

[4] T. F. Anderson, E. L. Wollman, and F. Jacob, "Sur les processus de conjugaison et de recombinaison génétique chez *E. coli III*. Aspects morphologiques en microscopie électronique," Ann. Inst. Pasteur (to be published).

[5] F. Jacob and E. L. Wollman, Symp. Soc. Exptl. Biol. **12**, 75 (1957).

[6] C. R. Fuerst and G. S. Stent, J. Gen. Physiol. **40**, 73 (1956).

[7] P. E. Hartman in *Genetic Studies with Bacteria*, Carnegie Institution of Washington Publication No. 612 (1956), p. 35.

[8] B. N. Ames in *Amino Acid Metabolism*, W. D. McElroy and B. Glass, editors (The Johns Hopkins Press, Baltimore, Maryland, 1955), p. 357.

[9] P. E. Hartman in *The Chemical Basis of Heredity*, W. D. McElroy and B. Glass, editors (The Johns Hopkins Press, Baltimore, Maryland, 1957), p. 408.

[10] C. Yanofsky, E. Lennox, and C. Richards, "Linkage relationships of the genes controlling tryptophan synthesis in *Escherichia coli*" (manuscript in preparation).

[11] M. Demerec and Z. E. Demerec, Brookhaven Symposia, Biol. No. 8 (1956), p. 75.

[12] S. Benzer and E. Freese, Proc. Natl. Acad. Sci. U. S. **44**, 112 (1958).

[13] D. B. Dunn and J. D. Smith, Nature **174**, 305 (1954).

[14] M. J. Bessman, I. R. Lehman, J. Adler, S. B. Zimmerman, E. S. Simms, and A. Kornberg, Proc. Natl. Acad. Sci. U. S. **44**, 633 (1958).

[15] R. M. Litman and A. B. Pardee, Nature **178**, 529 (1956).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

30
Coding Aspects of Protein Synthesis

CYRUS LEVINTHAL

Department of Biology, Massachusetts Institute of Technology, Cambridge 39, Massachusetts

TWO of the foregoing articles have dealt with certain aspects of formal genetics and others have dealt with the physicochemical studies of the structure of nucleic acids and protein. This article presents some of the current ideas about the relationship between genetics and the structure of these molecules. Many of these ideas are highly speculative and should be thought of only as current working hypotheses (or dogmas) which are being subjected to experimental tests.

Evidence indicating that genetic information can be contained in DNA was presented in previous articles and it is assumed then that there is a one-to-one correspondence between the position on the genetic map and the position along a DNA molecule. If DNA itself contains genetic information, it must be contained in the order of the bases along the two chains. But since only the base pairs adenine-thymine and guanine-cytosine are possible in the DNA molecule, the information is contained twice, once along each of the two chains. If the sequence of the bases along one chain is known, then that along the other chain is determined. This fact is, of course, the basis of the Watson and Crick hypothesis of the self-complementary nature of the duplication of the DNA molecule. However, for the moment, only the base sequence along one of the two chains is considered as being the genetic determinant.

From the fine-structure genetic analysis of Benzer,[1] Demerec,[2] and Pontecorvo,[3] it is known that functionally related mutations are near each other on the genetic map. As discussed by Lennox (p. 242), the cistron is a well-defined region of the map which is responsible for the control of a particular function. In interpreting this in chemical terms, Benzer made the hypothesis that a cistron corresponds to a length of DNA which controls the formation of a single polypeptide chain in a protein molecule. Both the cistron and the single polypeptide chain are operationally well-defined entities, and the definition of one does not depend upon the definition of the other. Thus, although this hypothesis is yet to be tested, it is subject to direct experimental verification.

There is another hypothesis[4] which is central to any discussion of the genetic control of protein structure. It states that the genetic information determines only the sequence of the amino acids within a protein and that the three-dimensional configuration of the formed protein molecule is a direct consequence of the amino-acid sequence. The main reason for accepting such a hypothesis is that the genetic map and the base sequence along the DNA molecule form one-dimensional structures, and apparently the only corresponding linear order in a protein molecule is the order of the amino acids along the polypeptide chains. It is evident that, in a formed protein molecule, the three-dimensional configuration is maintained by secondary bonds which exist between amino acids, linking either different polypeptide chains or widely spaced amino acids on the same chain. It is assumed, however, that the information specifying how those secondary bonds are formed is itself contained in the order of the amino acids. For example, covalent sulfur—sulfur bonds play an important role in maintaining the structure of many proteins, but it is assumed that their position in the molecule is determined primarily by the position along the peptide chain of the two cysteine molecules required for S—S bond formation. It should be stressed, however, that there is no direct evidence in support of this assumption, nor can a precise mechanism be formulated at present which might explain how the required bonds would be made. One only can speculate that the folding of a polypeptide chain which was formed along a linear template may require the sequential unpeeling of the chain from its template beginning from some specified place. One can imagine, too, that the correct folding could be influenced by some small molecule with which the finished protein would interact; as, for example, the inducer molecule in the case of induced enzyme synthesis. From the point of view of the over-all information transfer, such speculation as to the details of the mechanism is, of course, irrelevant, and is mentioned, at this point, only to indicate that the hypothesis is not entirely unreasonable.

In most of the cases where they have been studied, the active sites, or the business regions, of enzyme molecules are found to contain only a few amino acids. However, in order to carry out its enzymatic function, the amino acids in the active region of the enzyme must be arranged in a very precise, three-dimensional configuration. One can suppose that the information contained in the order of the *other* amino acids in the protein is necessary to specify the three-dimensional position of each of the amino acids in the active region relative to each other. Many of the *aa*'s coming from the active site may be necessary for the correct folding to occur. Once this has been accomplished, some of them could be eliminated without altering the active site. If this interpretation is correct, it follows that, in some portions of an enzyme molecule, one amino acid can replace another without producing any observable change in the enzymatic function of the molecule. For example, in some regions, the polypeptide chain may be folded in such a way that it forms an α-helix for a short region. Since most amino acids will do equally well as the com-

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

ponents of an α-helix, there could be many amino-acid replacements in such a region which would not affect the enzymatic function unless the particular amino acids were involved with secondary bonds to other regions of the molecule. On the other hand, it is likely that any amino-acid replacement in the active site of the molecule would cause a change in the enzymatic activity.

If it is assumed that the genetic information controls only the sequence of the amino acids along a polypeptide chain, then certain ideas suggest themselves which are independent of the detailed chemical mechanism by which proteins are synthesized. These arise from the fact that, although there are only four different kinds of bases in the DNA molecule, their sequence must uniquely determine the sequence of approximately twenty amino acids along the polypeptide chain. Thus, it would require at least three bases to select uniquely from among the twenty amino acids (4^2 is only 16 whereas 4^3 is 64). But the distance required for three nucleotide pairs along the DNA molecule is about 10 A whereas the distance between the amino acids on an extended polypeptide chain is only about 3.7 A. The first solution for getting around this apparent difficulty was suggested by Gamow.[5] He imagined a coding scheme in which n amino acids would be determined by a sequence of $n+2$ bases. Each amino acid would be determined by a sequence of three bases, but two adjacent amino acids would have two bases in common. Thus, two amino acids would be determined by a sequence of four bases; the first three members of the sequence, determining the first amino acid and the last three, determining the second amino acid. This type of overlapping code has two immediate consequences: First, there will be many mutations affecting two or even three amino acids; to date, no mutations of this type have been observed. Second—and this leads to a more serious objection to an overlapping code—not all pairs of amino acids can lie next to each other. In fact, the restrictions on the possible sequences of amino acids are so severe in the Gamow code that it seems to be impossible to use it to code some of the known amino-acid sequences. Brenner[6] has analyzed the known sequences of amino acids in proteins and has found that they are not consistent with *any* type of overlapping code in which two letters are used in common for adjacent amino acids. Brenner's analysis does rest on one further assumption, that the same code is used in all biological systems and for all proteins. This was a necessary assumption because, so far, only relatively small proteins have been analyzed as to the sequence of their amino acids and because it is not yet possible to obtain a sufficient number of adjacent pairs from any one protein. These two arguments, however, make it seem very unlikely that any variant of a double overlapping code is used in the synthesis of proteins.

Crick *et al.*[7] have avoided the spatial difficulty by making the chemically more plausible assumption that, when either DNA or RNA acts as a template for the formation of a protein, it is not the amino acid itself which fits on the template. They assumed that there could be some type of small adapter molecule which would attach to a specific amino acid on one end and to three nucleotides on the other. From what now is known of the chemistry of protein synthesis, one might suppose that the genetic information in the DNA is transferred first to a large RNA molecule which would act as the template for protein synthesis, and that the soluble RNA to which specific amino acids are attached could then be the adapter molecules. There is, as yet, no direct evidence that this model is correct, but it is interesting to note that the adapter hypothesis was made two years before there was any chemical information on the attachment of amino acids to soluble RNA. Having made this assumption, one is freed from the spatial argument and only need consider the problem from the point of view of coding. Crick *et al.*[7] have pointed out that the synthesis of proteins would be very much more efficient or, at least, more rapid, if each amino acid, presumably attached to its adapter, could place itself along the template independent of the other amino acids. Thus, a given position on the template should be such that it would accept a particular amino acid whether or not the adjacent spots were filled. In order to satisfy this condition and to use every three nucleotides to specify an amino acid, it is necessary to impose the condition that certain triplets (called *sense* triplets) of the nucleotides correspond to amino acids and that others (called *nonsense* triplets) do not. The code must be such that the six nucleotides required for two adjacent sense triplets should not contain any additional sense triplets among the overlap triplets. Thus, if both *xyz* and *uvw* correspond to amino acids, then *yzu*, *zuv*, *wxy*, and *vwx* must not correspond to an amino acid. Such a code is called commaless, since no additional information is required to indicate where amino acids are to be placed along the nucleotide sequence. In the language used by Elias (p. 221), this is a "jitter-free" code because it is independent of the ability to read correctly from one end as to the distance from that end.

Most of the sixty-four words which can be written with four letters taken three at a time can be eliminated by the comma-free condition. This set of three-letter words which can be used in a comma-free message is called a dictionary. The word "AAA" cannot be included in such a dictionary, because, if the word were repeated twice, the sequence "AAA" would appear in the overlap and the position of the amino acid would not be well defined. This condition eliminates the four possible words which have three identical letters and reduces the number of possible words in the dictionary from sixty-four to sixty. Furthermore, no cyclic permutation of a word can be in the dictionary. If "ABC" is included, then neither "BCA" nor "CAB" can be, because if either one of them were included and repeated

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

twice, "ABC" would occur in the overlap. Since, from any word, two others can be derived by cyclic permutation of the letters, the maximum size of the dictionary is reduced further from sixty to twenty.

Two twenty-word dictionaries which are essentially different from each other were presented by Crick *et al.*, and since no comma-free code can contain more than twenty words, this completes the proof that one is dealing with a maximal comma-free dictionary. Freudenthal[8] and Golomb *et al.*[9] have shown that a total of five, and only five, basic types of code can be constructed which satisfy the comma-free condition and contain twenty words (Fig. 1).

There is an additional difficulty with this type of code which is related to the fact that the information is contained twice in the DNA chain. Chemically, the two chains are antiparallel so that the over-all features of the molecule are invariant with respect to an end-to-end inversion. However, the informational content or the sequence of the nucleotide pairs clearly is not invariant with respect to this inversion. Or stating the problem differently, the information as to the protein to be formed by a given region of DNA will be different depending on which of the two chains is read. Delbrück[9] proposed that an additional requirement be added to the code such that, if the message were to read correctly along one chain, the complementary sequence which existed on the opposite chain would not contain any sense triplets. With this additional restriction, only ten words can be written with four letters taken three at a time, and it requires a sequence of four letters to supply enough information to specify twenty amino acids subject to this restriction. However, this additional requirement imposed by Delbrück is not absolutely necessary. One can, for instance, suppose that the information is transferred from a double-chain DNA molecule to a single-chain RNA molecule and that the information, specifying which of the two DNA chains will be used in forming the RNA molecule, is not present everywhere along the DNA, but only in short regions (for example, between cistrons). Thus, in the making of the RNA, one need not impose the commaless condition, whereas it still is required in the making of the protein from the RNA.

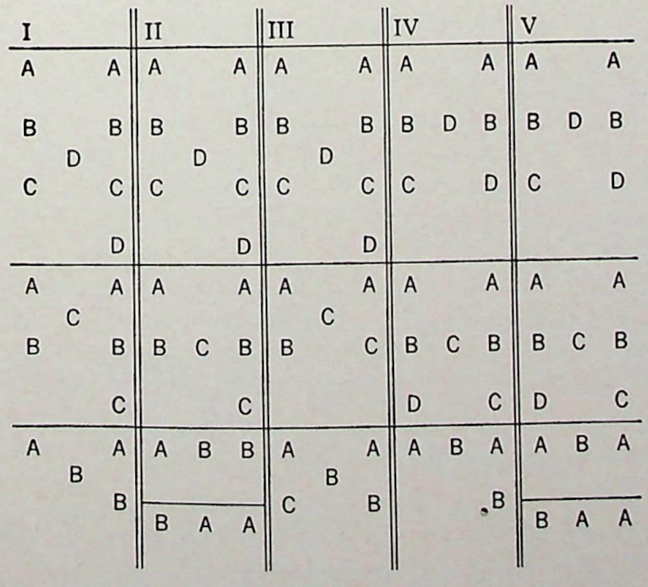


FIG. 1. The five possible types of commaless codes using four different letters taken three at a time (after Freudenthal[8]). In each group with a given middle letter, all combinations of the listed first and third letters are sense words.

In the absence of experimental data, it is clear that such speculation can go on more or less indefinitely, and that very elaborate codes can be imagined. For example, there is no particular reason to assume that the information specifying a given amino acid necessarily comes from contiguous nucleotides. One could imagine a mechanism in which every tenth nucleotide (that is, one every turn of the DNA molecule) would specify an amino acid so that one might be specified by nucleotides 1, 11, and 21 and the next by 2, 12, and 22, etc. At the moment, there is very little experimental evidence relevant to these questions. However, a great deal of work is being done in an attempt to obtain some data. In the hope of providing some further understanding of the current status of the field, the experiments are discussed in terms of the kind of question which might be asked and of the kind of information which, in principle at least, might be obtained.

It seems unlikely that, in the near future, it will be possible to analyze the sequence of the nucleotides in a piece of DNA which is responsible for a particular protein molecule. In addition to the difficulty of carrying out the sequence analysis of the bases along the DNA, there is another problem which seems even more formidable. Except in the case of some of the small viruses, the method is not clear for obtaining a preparation of DNA wherein each molecule of the material controls the same protein. And even in the smallest viruses, the amount of nucleic acid is very large for an analysis of this type. The work that is being done involves the analysis of how protein molecules are altered as a result of specific mutations. Most of the currently available experimental results relevant to this problem have come from the work on human hemoglobin done by Ingram.[10] He analyzed the hemoglobin molecule by treating it with the proteolytic enzyme, trypsin, and then subjecting the resultant digest to a two-dimensional paper electrophoresis and chromatography. Trypsin splits the polypeptide chain at every arginine and lysine residue. In the case of hemoglobin, this digestion yielded 28 different peptides. Hemoglobin made by different individuals can be compared by noting the various movements of different peptides on the paper. Sickle-cell anemia, a disease which is presumably due to a single mutation, results in an altered, but not totally inactive, hemoglobin. This alteration was shown to correspond to the change of a single amino acid. A total of six altered hemoglobins have been

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

examined by Ingram and, in three of these, it has been shown that only one amino acid is changed.[10] Two of the changes involve the same amino acid and one involves a change in a different region at the molecule. The other three alterations differ from these, but it has not been proved that only one amino acid is altered. In these studies with human hemoglobin, one selects for those mutations which produce a minor alteration in the functioning of the molecule. It is much less likely that an alteration which produces no physiological change would be examined. This is true also for any mutation resulting in the absence of a hemoglobin molecule which would appear as a recessive, lethal mutation, and, thus, would probably not be identified as affecting hemoglobin. However, it is important to note that no one of the cases so far examined represents changes in more than one of the peptides and of the three which have been totally characterized, none results in a change of more than one amino acid.

In several laboratories, work is directed toward obtaining a system with which fine-structure genetic analysis can be carried out and combined with chemical analysis of the protein controlled by the genetic region. In order to make such studies possible, one must work with a conditionally necessary protein whose absence prevents the organism from growing in some conditions but not in all. The substance made by the *rII* genetic region studied by Benzer[1] obviously satisfies this requirement. Although an intact protein is necessary if the virus is to grow in some bacterial hosts, it is not necessary for it to grow in other hosts. Likewise, there are many enzymes made by bacteria which are necessary for the growth of the bacteria only under certain environmental conditions. For example, fine-structure analysis has been done in bacteria by using the enzymes responsible for the synthesis of a certain amino acid. If the amino acid is supplied in the medium, the absence of the enzyme does not affect the growth of the organism. On the other hand, by placing the progeny organisms of crosses in a medium lacking the required amino acid, a very small number of recombinants can be selected which have the enzyme.

Under such conditions one can consider several classes of mutation. If the general concept of a commaless code is correct, then there will be many sequences of nucleotides which do not correspond to an amino acid. A mutation could occur which would result in such a nonsense sequence in the nucleotide chain. If "ABA" corresponded to a certain amino acid, then a mutation of the letter "B" to the letter "A" would result in the sequence "AAA" which is not in the dictionary. If this were to happen, it is possible that no protein molecule could be formed, since, presumably, the peptide chain no longer would be continuous at this point. This type of sense-to-nonsense mutation could result plausibly in the absence of an active protein, regardless of where it occurred in the molecule. In addition, there would be a class of sense-to-mis-sense mutations which could result in a totally inactive molecule. A replacement of valine by proline in an active site probably would result in a protein with no enzymatic activity. On the other hand, such a protein might still be formed and be recognized by its antigenic similarity to the normal enzyme. Also, the fact that the protein may be induced by the same inducing agent as caused the formation of the normal enzyme would be another factor in recognizing an alteration. Sense-to-mis-sense mutations are those which, in principle, may lead to a protein which could be analyzed chemically and whose difference from the normal enzyme could be determined. However, the sense-to-nonsense mutation would lead to the absence of a protein and, therefore, no information about the alteration of the amino-acid sequence. It may be possible, however, to obtain considerable information by studying the reversions to activity from the nonsense mutations. This is illustrated in Fig. 2(b) where the sequence equivalent to amino acid α first mutates so that it corresponds to a nonsense word. Then, reverse mutations are selected, some of which restore the amino acid α and others of which insert the amino acid β or γ. In this sort of mutation-reversion study, one expects to find, in some instances, that several different reversions could be obtained which produce an enzyme differing from the normal. However, each of the reversions would be expected to differ from the normal at the same position in the molecule. It would be possible to test this conclusion despite an inability to carry out fine-structure genetic analysis as long as the system is such that the reversions with some activity have a selective advan-

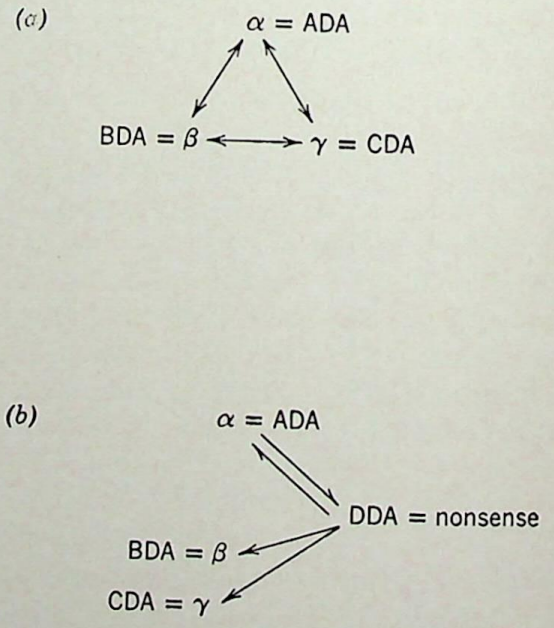


FIG. 2. (*a*) The conversion group α, β, and γ corresponds in all of the codes to changes in the first letter of the word, ADA. The amino acid α=ADA would belong to two other conversion groups also whose size depend on which code was used. (*b*) ADA could mutate to the nonsense word, DDA, also, and in this special case all of the possible reversions to sense words would be in the same conversion group. However, it is not generally true that the reversions will be in the same conversion group.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

tage. It has been shown[11] that, for all of the three-letter commaless codes, the number of different sense reversions from a nonsense mutation is at most four. This prediction could also be subject to experimental tests.

There are other classes of mutation which, although extremely useful in the genetic analysis, are not likely to be very helpful in this problem. These are the deletions or inversions which have been discussed earlier. In this case, one expects to find mutations which never revert to produce an active protein, and, as such, there could never be any analysis of an altered protein.

Another question which can be studied experimentally was mentioned earlier with reference to the work of Ingram.[10] It is concerned with whether or not a single mutation can affect more than one amino acid. In connection with this, mention should be made of one assumption which is implicit to all of these discussions. It is assumed that the probability of mutation is sufficiently low so that two letters in the nucleotide code will not be altered simultaneously. This assumption will be justified, if it is found that all mutations *do* affect a single amino acid only.

Another kind of analysis would become possible, if a sufficiently large number of transitions from one amino acid to another could be accumulated. Suppose [Fig. 2(a)] that a mutation occurs which carries amino acid α into β and another also occurs which carries α into γ. One can ask if β can mutate to γ. If it does, then the set of three amino acids is called a *conversion group*. A conversion set is defined as a set of amino acids, each member of which can be changed to every other member of the group by a single mutation. With the assumption that mutations change only a single nucleotide at a time, a conversion group must be produced by changing only one letter of the word which corresponds to an amino acid. Under these assumptions, the first prediction which would apply to any code is that no conversion group could contain more than four elements, corresponding to the four possible bases which could be inserted in a given position. If there are limitations on what alterations can occur in the nucleotide chain, the number of elements in a conversion group would be less. For example, if purines could be changed only to other purines and pyrimadines to other pyrimadines, no conversion group could contain more than two elements. A single amino acid could, of course, belong to several different conversion groups. The number of conversion groups to which a given amino acid belonged would be equal to the number of nucleotide letters which correspond to a single amino acid. Consequently, in any of the triplet codes, each amino acid would belong to three conversion groups. As can be seen from the five dictionaries indicated in Fig. 1, for the triplet commaless codes, some of the conversion groups would contain four elements, and others, three, two, or one. If a complete analysis of the conversion groups were possible, one could be led, unambiguously, to one of the five triplet codes. Conversely, if any code of more than three letters is written, the set of these conversion groups could be numerated and, therefore, in principle at least, any particular code could be tested with this type of analysis. However, in order to determine the conversion groups, a great deal of experimental data would be necessary.

All of the experiments mentioned so far require detailed chemical analysis of altered proteins but do not necessarily require fine-structure genetic analysis. However, if fine-structure genetic analysis can be combined with information as to the chemical alterations, two of the most fundamental questions can be investigated. The first involves the basic premise on which much of the speculation depends; that is, does the order and relative position of mutations along the genetic map correspond to the order and relative position of the corresponding altered amino acids along the polypeptide chain. If several different mutations can be mapped relative to each other and the corresponding altered proteins examined, one can investigate whether or not the relative positions are the same on the genetic map and on the polypeptide chain. One way of accomplishing this analysis would be to carry out the fine-structure mapping on the nonsense mutations which, since they produced no active protein, could be mapped with high resolution, and the amino-acid analysis on the reversions which have some activity and whose protein alteration could, in principle, be determined. Another problem which can be investigated, if the genetics could be combined with the chemical analysis, is to find whether or not the genetic determinants of a particular amino acid are adjacent to each other. If, for example, the information specifying one amino acid is interspersed along the nucleotide chain with the information controlling another amino acid, it would be evident from the finding that the mutants affecting one were interspersed among the mutants affecting another. In this way, the question as to whether or not the genetic code is truly a local code could be determined experimentally.

So far, this paper has dealt with much speculation, many questions, and a few experimental facts. The remainder of it is a brief review of the work going on in several laboratories which is directed toward obtaining some answers to questions raised. Ingram's studies with human hemoglobin are continuing, and the list of possible transitions from one amino acid to another is being extended. Even if the combined genetic and chemical analysis should become possible with proteins made by microorganisms, the work with hemoglobin will remain of prime importance because it can help to answer the question of whether or not the code is the same for different organisms.

Several systems are being investigated with microorganisms. The first is the enzyme trytophane synthetase which is being studied by Yanofsky. He has found[12] that this enzyme, under some conditions, can

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

catalyze two sequential reactions which ultimately lead to the formation of tryptophane. He has found that this enzyme can be altered by mutation so that either one of the two enzymatic activities can be eliminated, leaving the other intact. Those mutations affecting one of the activities are localized and are in a region of the genetic map which is adjacent to the region in which the second set of mutations can occur. In many of these cases, the bacteria form a material which is antigenically related to the enzyme produced by the normal cell. The effects of many of the mutations affecting this enzyme can be reversed by suppressor mutations which occur elsewhere on the genetic map in regions not adjacent to the original mutations.[13] This finding of unlinked suppressors is, at first sight at least, in complete contradiction to many of the assumptions used in this analysis. However, the detailed investigation of the chemical alterations associated with these mutations and suppressors has been undertaken only recently, and, until it is completed, several interpretations of the results are possible. Some of the mutations leading to a loss of enzymatic activity possibly may be owing to a certain alteration of the enzyme so that it can be inactivated by some normal cellular component. The suppressor mutations could then affect the amount of this component present in the cells. At least one such case, involving tryptophane synthetase, has been found by Suskind.[14] However, the resolution of this question must await further chemical analysis of the altered proteins.

The proteins which make up the *T*-even bacterial viruses have been investigated also,[15] with regard to finding a system for doing combined genetic and chemical studies. A great deal of genetic analysis has been carried out with these phages, and several proteins have been isolated from them. Unfortunately, the *rII* region, which Benzer has studied so extensively, does not control a protein which is contained in the phage itself. Rather, the control seems to be over one of the several proteins required in the infected cell for the formation of the phage. Two of the proteins which are in the phage particle have been purified and characterized, and, although differences can be observed between the strains, *T*2 and *T*4, it has not yet been possible to obtain different proteins by mutation of any *one* of the strains. There are, however, several types of mutations of these phages and it is very likely that some of them will be found to control these proteins.

In addition to the coding problem already discussed, another type of question can be approached by using the phage proteins. The membrane around the DNA of these viruses is composed of a large number of identical protein molecules. Mutants have been found which result in many of the formed phages having defective heads, so that it is possible that information can be obtained as to the organization of these small molecules into a large and relatively complex structure. Whether or not any progress toward the solution of this question is possible using the phage, it is clear that, in order to understand genetic control of an organism, it is necessary to know how organized structures, as well as protein molecules, are controlled.

Another attempt to obtain a system in which a conditionally necessary protein can be studied chemically and genetically is being undertaken by Garen and Levinthal using the bacterial enzyme, alkaline phosphatase. This enzyme splits the phosphate-ester linkage of organic phosphates and is necessary for the bacterial growth only if there is organic phosphate, but no inorganic phosphate, supplied in the medium. Alkaline phosphatase is made in high concentration by bacteria only if they are starved for phosphorus. Under these conditions, it is made in very large amounts. The ability to make a single enzyme in this high concentration seems to be a characteristic of bacteria, which, although they have the genetic capability of making thousands of different enzymes, also have additional mechanisms which govern the amount of each of these enzymes actually made under given culture conditions. Approximately 5% of the total cell protein occurs as this one enzyme. When the bacteria are producing alkaline phosphatase at the 5% level, it is comparatively simple to purify the protein. Mutants have been isolated after irradiation with ultraviolet light which showed no enzymatic activity, and others have been isolated which have reduced activity. Several of the latter type have been analyzed and found to produce an altered protein in the same high concentration which was observed in the "wild" type. From the mutants which lack any enzymatic activity, reversions have been obtained, some of which are like the "wild" type, and others, like the mutants with reduced activity. Fine-structure genetics can be done with these mutants, but as yet no information has been obtained concerning the alterations in the amino-acid sequence in the enzymes made by the mutant organisms.

It can be seen from this discussion that, although there is some evidence that a code exists which translates genetic information into amino-acid sequence, there is no information as to the *nature* of this code. On the other hand, a number of well-defined questions which can be answered experimentally have been formulated, and several different systems which are likely to yield some answers are being studied.

BIBLIOGRAPHY

[1] S. Benzer in *The Chemical Basis of Heredity*, W. D. McElroy and B. Glass, editors (The Johns Hopkins Press, Baltimore, Maryland, 1957), p. 70.

[2] M. Demerec, Z. Hartman, P. E. Hartman, T. Yura, J. S. Gots, H. Ozeki, and S. W. Glover, "Genetic studies with bacteria," Carnegie Institution of Washington, Publication No. 612 (1956).

[3] G. Pontecorvo, Cold Spring Harbor Symposia Quant. Biol. **21**, 171 (1956).

[4] F. H. C. Crick, Symp. Soc. Exptl. Biol. **12**, 138 (1958).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

[5] G. Gamow, A. Rich, and M. Ycas, Advances in Biol. and Med. Phys. 4, 23 (1956).
[6] S. Brenner, Proc. Natl. Acad. Sci. U. S. 43, 687 (1957).
[7] F. H. C. Crick, J. S. Griffith, and L. E. Orgel, Proc. Natl. Acad. Sci. U. S. 43, 416 (1957).
[8] H. Freudenthal, Koninkl. Ned. Akad. Wetenschap. Proc. A61, 253 (1958).
[9] S. W. Golomb, L. R. Welch, and M. Delbrück, Biol. medd. Dan. Vid. Selsk. 23, 9 (1958).
[10] V. M. Ingram, Nature 180, 326 (1957), and personal communication.
[11] I. Tessman (personal communication, 1958).
[12] C. Yanofsky and J. Stadler, Proc. Natl. Acad. Sci. U. S. 44, 245 (1958).
[13] C. Yanofsky, Science 128, 843 (1958).
[14] S. R. Suskind and L. I. Kurek, Science 126, 1068 (1957).
[15] S. Brenner, S. Benzer, S. Champe, and G. Streissinger (personal communication, 1958).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

REVIEWS OF MODERN PHYSICS VOLUME 31, NUMBER 1 JANUARY, 1959

31
Induced Enzyme Synthesis

DAVID S. HOGNESS

Department of Microbiology, Washington University School of Medicine, St. Louis 10, Missouri

I. INTRODUCTION

A USEFUL hypothesis of modern genetics and that forming the basis of the preceding papers is that the genes (specifically, the cistrons) within a cell determine which proteins among all possible proteins a given cell can conceivably synthesize. Such genes are thus considered as the primary determinants of a given cell's allowance of proteins.

There are, however, other determinants of this allowance which, although secondary to genetic control, are of interest in any consideration of the mechanism of protein synthesis in that they specifically affect the synthesis of a single protein. Three phenomena have been sufficiently delineated to exemplify such secondary specific determinants. These are:

(1) Specific antibody synthesis resulting from exposure of cells to a given antigen.
(2) Induced enzyme synthesis.
(3) Repression of enzyme synthesis.

This paper is concerned solely with the last two phenomena, which quite possibly represent two forms of the same basic event.

Induced enzyme synthesis can formally be defined as the increase in the ratio of the rate of synthesis of a given enzyme to the rate of synthesis of total cell protein resulting from exposure of cells to compounds (inducers) which are identical or structurally related to the substrates of the given enzyme.[1] The vast majority of known instances of induced enzyme synthesis is derived from microorganisms, particularly the bacteria. This disproportionate representation is probably not real, for we have some reason to suspect that enzyme induction is operative in most cell types. Rather, it would seem that our observations are selected by the experimental techniques available for the manipulation of bacterial and fungal cell populations which have not, until very recently, been available for the culture of other cell types. A second consequence of this precision of bacterial manipulation is that the best-characterized and interpretable induced enzyme systems are found in the bacteria.

Induced enzyme synthesis in bacterial populations has been known for approximately seventy years, although, during a majority of its observed lifetime, this phenomenon has been shrouded in a teleological disguise by being named "enzymatic adaptation." It is only in the past decade that quantitative studies have allowed one to define this phenomenon as an induced enzyme synthesis not necessarily related to any increase in fitness of the cell in which it occurs.

Even in bacterial populations in which there are the most numerous examples of enzyme induction, it is quite clear that the majority of known enzyme-forming systems cannot be classified as inducible in that the enzymes are formed at considerable rates in the absence of exogenous inducers, and this rate of synthesis cannot be specifically increased by exposure of cells to their substrates or such structural analogs of these substrates as have been tested. Such enzymes are often referred to as constitutive enzymes in order to differentiate them from the induced variety. It must be stressed that the terms induced and constitutive do not describe the properties of an enzyme *per se*, but rather describe the properties of an enzyme-forming system. Thus, the same protein molecule can result from an induced enzyme synthesis in one bacterial population while being the resultant of constitutive enzyme synthesis in another, genetically distinct, bacterial population.

There is then an apparent dichotomy among enzyme-forming systems, being either inducible or constitutive. While at present the data are not sufficient for a unique explanation of this dichotomy, an attempt is made here to correlate these two systems to a common working hypothesis, thereby making the questions to be asked more specific and, it is hoped, experimentally answerable. It would not be profitable for the purpose of this paper to review the many known systems of enzyme induction or, in fact, even one of these systems in all of its detail. (For this purpose, see references 2–11). Rather, emphasis is placed on the salient features of one system that are relevant to a precise experimental illustration of the definition of induced enzyme synthesis given above and to the apparent dichotomy between induced and constitutive synthesis. The system of choice for this purpose is the β-galactosidase of *E. coli*, since it has been analyzed in perhaps the most detail as regards the enzyme protein itself, its induced synthesis, and the genetic and physiological relationship between its induced and constitutive synthesis.

II. β-GALACTOSIDASE INDUCTION

(A) Characteristics of the Enzyme Protein[8]

Since studies of the induction process depend upon the measurement of enzyme activity as a measure of the amount of enzyme protein present at any given time, it is obviously essential that a direct correlation between these two quantities can be made for the variety of conditions employed. This, in turn, demands that the catalytic and structural parameters of the enzyme in question be sufficiently determined. The β-

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

galactosidase of *E. coli* has been subjected to such determinations and it is useful in the discussion of its induced synthesis if some of its basic properties are briefly described.

(a) The characteristic reaction catalyzed by β-galactosidase is represented in Fig. 1, namely, the hydrolysis of β-D-galactosides. Many glycosides have been tested as substrates or competitive inhibitors of this enzyme yielding the conclusion that the minimum requirement for affinity for the active site on the enzyme is the existence of the β-D-galactopyranosidic ring.[8] Whereas all of the β-D-galactosides tested are substrates, an interesting class of compounds, the β-D-thiogalactosides in which sulfur replaces the oxygen atom of the galactosidic linkage, function only as competitive inhibitors. The interest in these compounds lies in their capacity to function as inducers of β-galactosidase synthesis, without at the same time being hydrolyzed by the enzyme whose synthesis they induce.

(b) Purified preparations of β-galactosidase have been obtained which contain at most 1 to 2% contaminating protein.[8] This allows a direct correlation of the unit of catalytic activity* with the mass of enzyme, yielding the value of 1 catalytic unit per 3.0×10^{-9} g of protein. Since Cohn[8] has determined by equilibrium-dialysis methods that there is one active site per molecular weight 1.3×10^{5}, activity measurements can yield a determinination of the number of active sites. Thus, one catalytic unit equals 1.4×10^{10} such sites.

(B) Induction Phenomenon

A culture of *E. coli* growing in a medium of inorganic salts with a nongalactosidic carbon source, such as succinic acid, produces only trace amounts of β-galactosidase. The addition of a suitable galactoside to such a growing culture is immediately followed by a sharp increase of over 1000-fold in the rate of synthesis of this enzyme. This high rate of synthesis is maintained as long as the bacteria grow in the presence of the inducing galactoside (inducer). However, if the inducer is removed, the rate of synthesis falls directly to the original small value. The quantitative aspects of this situation are presented in Fig. 2. As shown in the left-hand graph of this figure, from 0 to 120 min the bacteria are growing in the absence of the inducer and the amount of enzyme per unit weight of bacteria is quite low, approximately 7 units per mg dry weight of bacteria, which is equivalent to 20 active sites per bacterium. However, immediately upon adding the inducer, one sees that the amount of enzyme in the culture increases rapidly against a background of constant growth rate. Thus, while the bacteria have only doubled in amount, the β-galactosidase activity has increased by a factor of

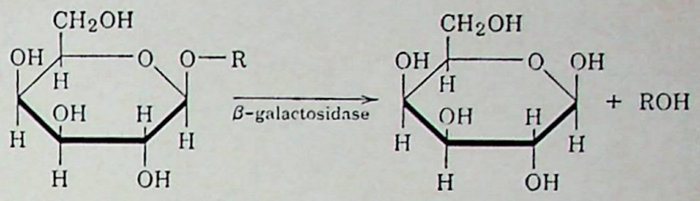


FIG. 1. Hydrolysis of a β-D-galactoside catalyzed by β-galactosidase.

approximately 1200. Plotted in this way, one observes the transition of the culture from one steady-state condition in which the inducer is absent (noninduced state) to a second steady-state condition with inducer present (fully induced state). The bacteria in the fully induced state contain 8.5×10^{3} catalytic units per mg dry weight of bacteria or approximately 24×10^{3} active sites per bacterium. This increase of about 1200-fold in the amount of enzyme per bacterium in the fully induced state over that in the noninduced state means that the rate of synthesis of β-galactosidase relative to the rate of synthesis of bacterial mass (i.e., the differential rate of β-galactosidase synthesis[12]) has increased via the induction process by the same factor of 1200.

It is this differential rate of enzyme synthesis which is of particular interest, since it is a more direct measure of the specific effect of the inducer on the rate of enzyme synthesis. One, therefore, wants to ask the question: how does this differential rate of β-galactosidase synthesis change during the induction process? This question is best answered by plotting the β-galactosidase activity of the culture *vs* the bacterial mass, a procedure which yields the right-hand graph of Fig. 2. Here, it is seen that the differential rate of β-galactosidase synthesis (i.e., the slope of the curve, *P*) changes from that of the noninduced state (7 units per mg dry weight) to that of the induced state (8.5×10^{3} units per mg dry weight) almost immediately after the introduction of the inducer, the transition taking place in less than 2 min, which is the minimum time detectable by the experimental techniques employed.

This high differential rate of synthesis remains constant as long as the inducer is present in the medium. However, as is shown in both graphs of Fig. 2, the inducer effect is readily reversible, since removal of the inducer immediately restores the differential rate of synthesis characteristic of the noninduced state, again with no appreciable change in the growth rate. Thus, in the case of β-galactosidase induction, it would appear that one has a system in which the synthesis of a given protein can be initiated or stopped by the simple addition or removal of a compound which, because it cannot function as a carbon or energy source, apparently does not influence the rate of synthesis of the vast majority of other cell constituents.

(C) Enzyme Induction as *de novo* Synthesis

Consider what is meant by synthesis as used in the preceding paragraphs. While it is true that the physical

* One unit of catalytic activity is defined as that amount of enzyme which will cause the hydrolysis of *o*-nitrophenyl-β-D-galactoside to occur at the rate of 1 mμmole/min at 28°C and *p*H 7.1 in $1.0\times10^{-1}M$ sodium phosphate and $2.7\times10^{-3}M$ *o*-nitrophenyl-β-D-galactoside.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

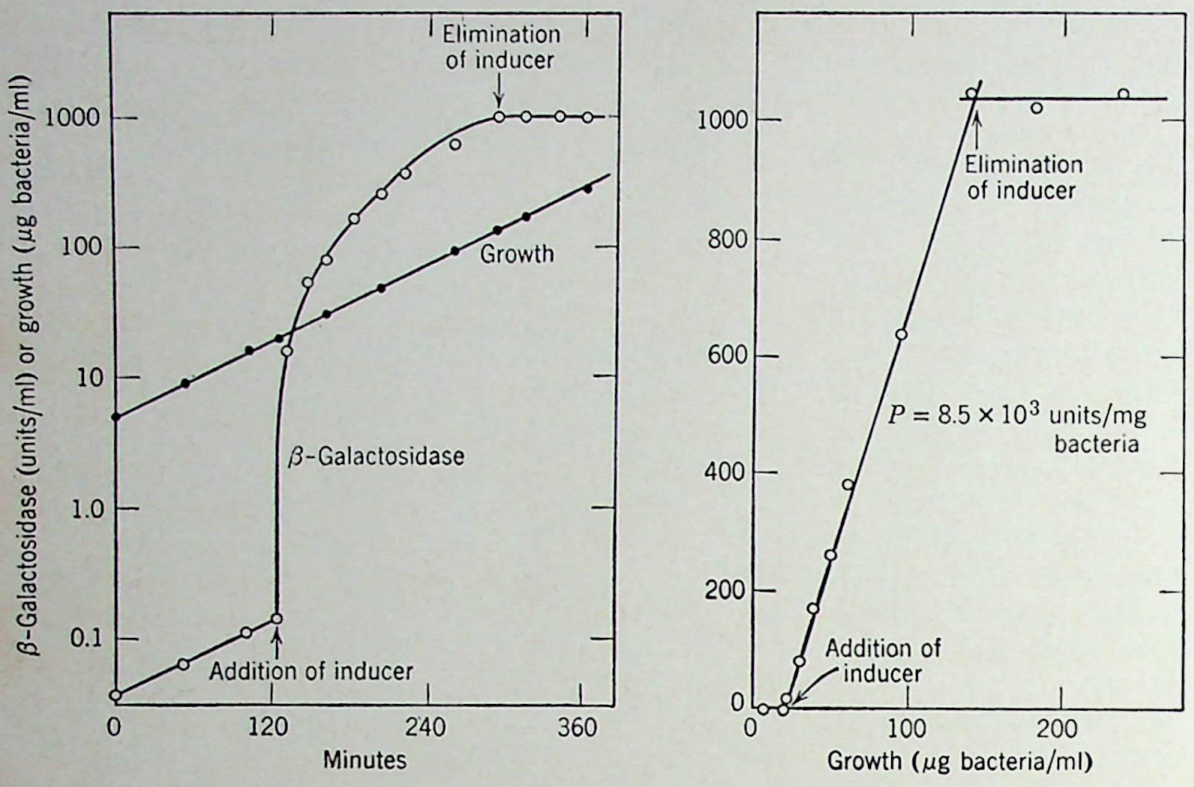


FIG. 2. Kinetics of β-galactosidase induction in *E. coli* ML 3. Inducer is isopropyl-β-D-thiogalactoside at 1.0×10^{-3} *M*.

and catalytic parameters of the active enzyme are sufficiently well known that one can calculate the mass of the active enzyme from the measurements of the catalytic activity present in the culture, it does not necessarily follow from this knowledge alone that an increase in such activity (and, therefore, of the mass of active enzyme) results from the *de novo* synthesis of the active enzyme. Within the context of the information so far presented, it is conceivable that the increase in active enzyme involved in the induction process results from the formation of a new protein structure (active enzyme) derived from inactive proteins present in the noninduced state. This transformation from inactive to active material could result (a) from a zymogen- to enzyme-type reaction typical of protease activations, (b) from the formation of a complex active protein made up of inactive subunits, or (c) from the addition of a few amino acids in peptide linkage to an inactive protein. In all of these cases, the induction process would have to be considered as independent of *de novo* synthesis in the sense that protein precursors of active β-galactosidase would be present in the noninduced state, and, consequently, that the total protein potentially or actually available as active enzyme would not necessarily be changed as a result of induction. It is, therefore, critical that one determine whether the induction phenomenon corresponds to the activation of protein precursors or to *de novo* synthesis, before it is possible to assess the value of this phenomenon as a tool for the study of general mechanisms of protein synthesis.

Although, by 1954, there were several lines of evidence which led, rather indirectly, to the tentative conclusion that β-galactosidase induction did indeed correspond to the *de novo* synthesis of this enzyme,[12] it was not until that year that conclusive evidence was available for the proof of this correspondence.[13,14] The direct proof of this correspondence came from experiments carried out in collaboration with Cohn and Monod[13], and simultaneously from the work of Rotman and Speigelman.[14]

The experiments consist of determining whether or not proteins present in noninduced cells are ever incorporated in or associated with the induced β-galactosidase protein by simply labeling the proteins in the noninduced cells with a suitable radioactive isotope and measuring the amount of radioactivity present in the enzyme when induced in nonradioactive media. Thus, in our experiments, *E. coli* were first grown in a simple salts medium in which the sole source of sulfur was S^{35} labeled sulfate. The amount of sulfate was so adjusted relative to the other components of the media that the bacteria stopped growing for lack of sulfate. In this starved condition, all of the radioactivity associated with the cells is contained within the trichloracetic-acid precipitable protein, and since no inducer was added to the medium, we have the condition of a unique radioactive labeling of the proteins of noninduced cells. This constitutes Step I (Table I) of the experiment.

In Step II, nonradioactive sulfate is added to the medium simultaneously with the inducer, methyl-β-D-thiogalactoside. This allows growth and induction to occur simultaneously. When the mass of the cells had increased so that 16% (aliquot *A*), 28% (aliquot *B*), and 43% (aliquot *C*) of the total growth had occurred in the nonradioactive inducing medium, aliquots of the bacterial population were removed, extracted, and, after purification of the β-galactosidase in these extracts, the amount of radioactivity associated with the enzyme determined.

For this experiment to succeed, it was necessary to develop purification techniques for the isolation of the

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

enzyme from less than one gram of cells of which only 0.16% was β-galactosidase. That this was possible is demonstrated by the data in Table I relevant to the isolation control in which unlabeled enzyme was mixed with a labeled extract of noninduced cells, such that the ratio of radioactivity to enzyme activity (K) in the mixed extract was the same as that for the crude extract of aliquot A. By comparing the K-ratio of the enzyme isolated from this reconstruction mixture with that isolated from bacteria grown and induced in the S^{35} labeled medium (fully labeled enzyme control), it is clear that the purification procedure yields an enzyme preparation which contains less than one percent bacterial protein contamination (0.4% in this case).

The enzyme purified from aliquots A, B, and C contained, respectively, 0.1, 0.8, and 0.1% of the radioactivity associated with the fully labeled enzyme. Since this is within the limits of the purification procedure as indicated by the isolation control data, and since these values do not progressively decrease with increase in the fraction of total growth occurring in Step II, it was concluded that less than 0.8% of the sulfur of the enzyme formed in Step II was derived from proteins present in the noninduced state (Step I). This result, taken in conjunction with the fact that, in aliquot A, the enzyme level was only 5% of that found in fully induced bacteria, indicates that if any protein precursor of β-galactosidase exists in the noninduced bacteria, its level must be less than 0.04% of that for β-galactosidase in fully induced bacteria or less than 10 precursor molecules of molecular weight 1.3×10^5 (i.e., the molecular weight per active site of β-galactosidase) per bacterium, assuming of course that such precursors contain the same percent sulfur as does β-galactosidase. This conclusion is enforced by the fact that Rotman and Spiegelman[14] found essentially the same results using C^{14} as the labeling material.

The results of these experiments allow one to conclude that, if there are any protein precursors existent in noninduced cells, they cannot form an appreciable contribution to the increase in active enzyme found upon the addition of the inducer, since it would take only a few seconds to form 10 new active sites per bacterium under the conditions of induction indicated in Fig. 2. Thus, when cells are exposed to inducer molecules the over-all rate of conversion of amino acids into a specific protein, β-galactosidase, is drastically and immediately increased, and consequently induced enzyme formation is equivalent to induced *de novo* synthesis. It would appear then, that the induction phenomenon should be accounted for in any working hypothesis that attempts an explanation of the general mechanism of protein synthesis.

TABLE I. Incorporation of sulfur into β-galactosidase synthesized by labeled *E. coli* cells in nonradioactive medium.

Experiment	Percent maximal enzyme level	$K = \frac{\text{radioactivity}}{\text{enzyme activity}}$	Percent of K for fully labeled enzyme[a]
Step I	0.06	(0.45)[b]	(100)
Step II—*A*	4.8	0.0050	0.1
B	32	0.0043	0.8
C	58	0.00072	0.1
Controls			
Fully labeled enzyme	100	0.45	100
Isolation control	(4.8)[c]	0.0018	0.4

[a] Corrected for basal activity.
[b] ( ) Basal level assumed to be equal to fully labeled enzyme.
[c] Reconstruction of Extract *A* by mixing unlabeled enzyme with a labeled extract of noninduced cells.

Two further conclusions resulted from our labeling experiments. The first stems from the preceding experiment and states that, in exponentially growing *E. coli* cells, the vast majority of non-β-galactosidase proteins are stable in the sense that they are not broken down to their constituent amino acids at rates that are appreciable in relation to the rates of synthesis. If the state of proteins within these cells consists of a continual synthesis from and breakdown to their constituent amino acids (i.e., state of "dynamic equilibrium"), then one would expect the β-galactosidase synthesized in Step II of the above experiment to be labeled with S^{35} as a result of the breakdown of the radioactive proteins. A simple calculation from the data of this experiment yields the conclusion that the average rate of breakdown of the non-β-galactosidase proteins must be less than one percent of their average rate of synthesis.

It is, therefore, not surprising to recall that, when the inducer is removed from an exponentially growing culture of *E. coli* cells (Fig. 2), the β-galactosidase formed while inducer was present remains constant, simply being diluted out in the exponentially increasing population. This would indicate that β-galactosidase is also stable once formed. Because of the possibility that this constancy of enzyme activity results from an equal rate of synthesis and degradation of β-galactosidase, and to test whether or not β-galactosidase is also stable in the presence of the inducer, we carried out further S^{35} labeling experiments.[13] These experiments yielded the expected conclusion that the β-galactosidase molecules in exponentially growing cells are stable in both the absence and the presence of inducers. Thus, it can be concluded that the induced *de novo* synthesis of β-galactosidase is, like the synthesis of other proteins in *E. coli*, essentially an irreversible process, and the so-called "dynamic state" is not a concept that need be involved in an explanation of the mechanism of such synthesis.

(D) Induction at the Cellular Level

While it is clear from the foregoing discussion that one may interpret the kinetics of formation of β-galactosidase activity during the induction of a bacterial population as the kinetics of *de novo* synthesis of this enzyme in that population, one can only extrapolate these kinetics to the cellular level if one makes the as-

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

sumption that all cells participate equally and simultaneously in this induced synthesis. Even though one can demonstrate a high degree of genetic homogeneity for the bacterial population being induced, there is no *a priori* promise that this assumption of equal and simultaneous participation of all cells is justified. It is, therefore, critical to the correct interpretation of the kinetics of induced enzyme synthesis in cultures that one have a knowledge of the distribution of rates of induced synthesis among the cells of the bacterial population.

There are two known types of conditions that can cause a heterogeneous response of a bacterial population when exposed to inducer molecules. The simplest of these types is a condition whereby the catalytic activity of the first enzyme molecules synthesized by a given cell increases the probability of synthesis of future enzyme molecules by that cell. As an example of this condition, consider a culture of *E. coli* whose growth has stopped because of exhaustion of glucose in the medium. If lactose (4-glucose-β-D-galactoside) is added to such a culture, it will have two functions: to induce the synthesis of β-galactosidase and, as a substrate of this enzyme, to provide the only available carbon and energy source necessary for cell synthesis. Under these circumstances, the induced synthesis of the first enzyme molecules must depend upon traces of enzyme already present or upon internal bacterial reserves. Since synthesis of the first enzyme molecules by any cell will increase the availability of carbon and energy for further synthesis by that cell, the cells which have a head start will increase their advantage. Thus, any initial heterogeneity in the population with regard to internal reserves or initial amount of β-galactosidase would be expected to result in an exaggerated heterogeneity in respect to enzyme content during the initial stages of growth on lactose. This expectation has been verified experimentally by Benzer.[15]

Elimination of this factor favoring heterogeneous response is quite simple and consists of using an inducer which is not a substrate of β-galactosidase (e.g., the β-D-thiogalactosides), or alternatively of employing mutant bacteria which cannot further metabolize the products of hydrolysis of the inducer, while at the same time providing a nongalactosidic carbon and energy source (e.g., succinic or lactic acid) for cell synthesis. These conditions have been defined as "conditions of gratuity" by Monod and Cohn[3] and are the conditions employed in the experiment illustrated in Fig. 2. Benzer[15] found that, in the induced synthesis of β-galactosidase under conditions of gratuity and at saturating levels of inducer concentration, the bacterial population exhibits a homogeneous reaction in that essentially all cells participate to the same degree. Hence, under these conditions, the kinetics of induction of a culture represent the kinetics of induction of single cells.

Although most readers of Benzer's work concluded that conditions of gratuity were a sufficient guarantee for the assumption of population homogeneity during induction, Benzer himself cautioned against applying this assumption to conditions of less than saturating concentrations of inducer even though the induction was "gratuitous." Indeed, a second more subtle factor causing heterogeneous response at less than saturating concentrations of inducer has been discovered by Rickenberg, Cohen, Buttin, and Monod[16] (also referred to in reference 9). Thus, these workers found that, in wild-type *E. coli*, the β-galactosidase induction can be divided into two separate processes. The first of these involves an active transport of the inducer into the cell such that inducer concentrations inside the cell are much greater than those in the medium. The second is the actual induction of β-galactosidase synthesis inside the cell at a rate determined by the internal concentration of inducer. The active transport of inducer is accomplished by a unit, called galactoside-permease, which has been demonstrated to have most of the characteristics usually associated with enzymes and whose synthesis is induced by many of the same compounds which induce β-galactosidase synthesis, again at a rate determined by the internal concentration of inducer.

This transport mechanism obviously introduces another factor that would be predicted to encourage a heterogeneous response of a population upon exposure to inducer, providing the probability that a given cell will produce its first permease molecule per smallest detectable time unit is small. Thus, those cells which do synthesize their first permease unit will have an increased internal concentration of inducer which will, in turn, increase the probability of synthesizing the second permease unit above that for the first. Thus, those cells which by chance are the first to synthesize permease units will increase their advantage and rapidly pass to the stage of saturating internal concentrations of inducer and maximum rate of synthesis of both the permease and β-galactosidase. The distribution of rate of enzyme synthesis in the bacterial population at any given time after addition of the inducer will thus depend upon the probability that a given cell will synthesize its first permease unit and upon the time interval between this event and when maximum rate of synthesis is reached.

Novick and Weiner[17] and Cohn[8] have analyzed the kinetics of β-galactosidase induction in the wild-type *E. coli*, which is inducible for both the galactoside-permease and β-galactosidase. They have found that these kinetics are consistent with the foregoing theory with the addition that the probability of synthesis of the first permease units is determined by the inducer concentration, and that the time interval between this event and when maximum synthetic rate is achieved is very small, even at the lowest inducer concentrations studied. This means that, at inducer concentrations well below saturation, the distribution of rates of enzyme synthesis at any given time is essentially an all-or-none distribution; i.e., cells are synthesizing enzyme either at the maximum rate or at the minimum noninduced rate.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

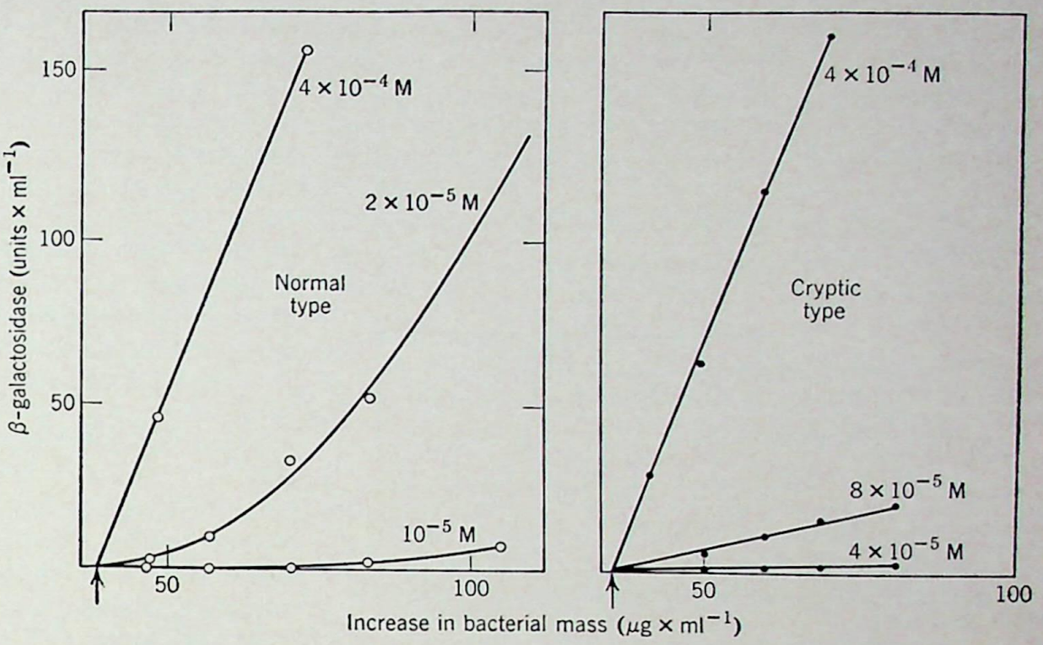


FIG. 3. Kinetics of β-galactosidase induction in normal *E. coli* and in the cryptic mutant. Inducer is isopropyl-β-D-thiogalactoside at the concentrations indicated above [from L. A. Herzenberg, "Studies on the induction of β-galactosidase in a cryptic strain of *Escherichia coli*," Biochim. et Biophys. Acta (to be published)].

As this induction at low inducer concentrations continues, a larger and larger fraction of the bacterial population is synthesizing enzyme at the maximum rate, and, consequently, a plot of enzyme activity *vs* increase in bacterial mass during the induction of such a culture will yield a curve with a slope (differential rate of synthesis) that increases with increase in growth to some final constant value (left-hand graph of Fig. 3). As the external inducer concentration is raised, the probability of synthesis of the first permease unit increases, and, consequently, the time duration of the phase of increase in differential rate of synthesis, or what may be termed the heterogeneous phase, decreases until it is no longer detectable. Under these conditions of saturating inducer concentration, for all intents and purposes, the distribution of rates of β-galactosidase synthesis in the population at any given time after addition of inducer is uniform, and, consequently, the kinetics of induction of the culture represents the kinetics of the individual cells.

However, it is quite clear that wild-type *E. coli*, in which both the galactoside-permease and the β-galactosidase are inducible, cannot be used to define the kinetics of induction of single cells at less than saturating concentrations of inducer. This could be accomplished if the galactoside-permease could be eliminated, thereby removing the last contributing factor toward heterogeneous response. The most convenient and complete method of such elimination is by the isolation and use of *E. coli* mutants which have lost the ability to form galactoside-permease, but retain the property of being inducible for β-galactosidase. Such mutants have been isolated and are commonly called inducible cryptics, in that, while β-galactosidase can be induced in these strains, because they lack the permease, the induced enzyme is essentially hidden from external substrate (e.g., lactose), the rate of hydrolysis in whole cells being limited by passive diffusion of the substrate.

One of these inducible cryptics (*E. coli* ML3) has been used extensively by Herzenberg[18] to define the kinetics of induction throughout the range of inducer concentrations that produce measurable increases in the rate of β-galactosidase synthesis. It is also the strain used in the experiment depicted in Fig. 2. An interesting result of the experiments of Herzenberg is that, for each inducer concentration tested, the differential rate of synthesis remains constant from the time of addition of the inducer, although the actual value of this rate is determined by the inducer concentration (Fig. 3). Under the assumption that, by employing conditions of gratuity and the inducible cryptic strains of *E. coli*, all factors encouraging a heterogeneous response during induction have been eliminated, one can interpret these kinetics of constant differential rate of synthesis for the culture as representing the kinetics of induction of individual cells. It must, however, be noted that, while this somewhat tortuous path of discovery and elimination of the sources of heterogeneous response during the induction process in *E. coli* wild-type makes this assumption quite reasonable, there is at present no direct experimental evidence available that completely justifies it.

If one admits this assumption, then the kinetics of constant differential rate of β-galactosidase synthesis during induction of the inducible cryptic strains (Fig. 3) allows one to conclude that the number of enzyme-forming units per cell that are activable by the inducer remains constant during induction. For, if the number of these units per cell should change during induction, then, in the presence of a constant subsaturating internal concentration of inducer, the differential rate of synthesis should change proportionately. The precision of the conclusion as to the constancy of β-galactosidase

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

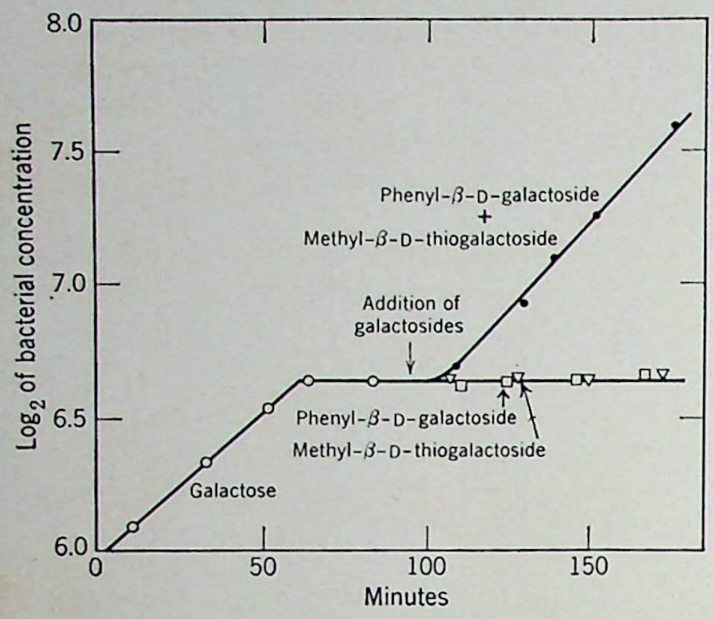


FIG. 4. Growth of *E. coli* ML 30 on galactose and the indicated galactosides demonstrating the independence of inducer and substrate functions.

forming units per cell is equivalent to the precision of the experimental observation of constant differential rate of synthesis. Consequently, if any change in the number of these units occurs during induction, it must occur in every cell of the population within two minutes of the moment the inducer is added. Since this seems extremely unlikely, it can be inferred that the same number of enzyme-forming units exist in noninduced cells as in induced cells, these units being active only in the presence of the inducer.

(E) Independence of Substrate and Inducer Functions

A corollary to this conclusion is that the catalytic activity of β-galactosidase is not functional in its own synthesis; i.e., the process of enzyme induction is independent of enzyme action. The kinetic derivation of this postulate is confirmed by the observation that the property of being a substrate for β-galactosidase is neither a sufficient nor necessary condition for a compound to function as an inducer. Thus, phenyl-β-D-galactoside is an excellent substrate of β-galactosidase, but is not functional as an inducer. Methyl-β-D-thiogalactoside, on the other hand, is not a substrate of the enzyme, but is an excellent inducer. The inverse relationship of these two compounds is demonstrated quite clearly by a simple growth experiment shown in Fig. 4. An inducible strain of *E. coli* was grown in a medium containing galactose as the sole carbon source until growth stopped (*ca* 60 min) as a result of depletion of galactose. Since galactose does not induce β-galactosidase synthesis, these starved cells contain only trace amounts of the enzyme. After waiting a short time to insure complete starvation, the starved cells were divided into three parts, and the galactosides indicated in Fig. 4 were added as the sole carbon source. The cells did not grow in the presence of methyl-β-D-thiogalactoside because, although it can induce β-galactosidase synthesis, it is not a substrate and, consequently, cannot function as the necessary carbon source. No growth is observed in the presence of phenyl-β-D-galactoside, because it is not an inducer of the enzyme necessary for it to function as a carbon source. However, when the two galactosides are added together, growth occurs since methyl-β-D-thiogalactoside functions as the inducer and phenyl-β-D-galactoside serves as the substrate, yielding galactose, the necessary carbon source.

In addition to the qualitative data as to independence of substrate and inducer functions, quantitative measurements of the ability of many thiogalactosides to induce β-galactosidase synthesis and the ability to complex with the active site of this enzyme show no correlation. Thus, phenyl ethyl-β-D-thiogalactoside is an extremely good competitive inhibitor of β-galactosidase action ($K_I = 1.5 \times 10^{-5} M$), but can induce the synthesis of β-galactosidase to a rate that is only one-eightieth that obtainable with methyl-β-D-thiogalactoside, although this latter compound is about 800-fold less efficient as a competitive inhibitor ($K_I = 1.2 \times 10^{-2} M$). On the other hand, melibiose, an α-galactoside which is neither a substrate nor an effective competitive inhibitor of β-galactosidase, is quite effective as an inducer of this enzyme. Thus, it would appear that, in activating the enzyme forming units within the cell, the inducer reacts with some material having a different specificity than that associated with the active site of β-galactosidase.

There remains, however, one minimum structural requirement common both to inducers and to substrates or competitive inhibitors—namely, that they contain the galactopyranosidic ring structure.† The inference, then, is that the site at which the inducer reacts to activate the enzyme-forming unit is structurally similar but not identical to the active site of β-galactosidase.

In concluding this experimental definition of induced enzyme synthesis, the main conclusions that have been obtained are listed below.

1. The induction process involves the complete *de novo* synthesis of the enzyme from its constituent amino acids.

2. Induced enzyme synthesis is a virtually irreversible process, the enzyme being stable in the presence or absence of inducer.

3. The number of enzyme-forming units per cell remains constant during the induction process; that is, the inducer does not change the rate of synthesis of enzyme-forming units, but simply activates such units.

4. The site at which the inducer reacts to activate the enzyme-forming unit is structually similar but not identical to the active site on the enzyme.

---

† This is strictly true only if the *C*6 carbon is not considered part of the ring, since the α-L-arabinosides (which are derivatives of β-D-galactosides lacking *C*6) exhibit the property of being weak substrates for β-galactosidase.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

In addition to the above conclusions, one should mention another of equal importance but which, at present, cannot be derived from data on β-galactosidase induction. It is that the inducer acts as a catalyst in the sense that one molecule of inducer may cause the formation of more than one molecule of enzyme. This observation comes from experiments concerning the penicillin-induced synthesis of the enzyme penicillinase in *Bacillus cereus*.[11] It should be noted that each of the foregoing conclusions has been experimentally confirmed in only one, or, at the most, a very few systems and that, consequently, their generality remains to be shown.

III. CONSTITUTIVE SYNTHESIS OF β-GALACTOSIDASE

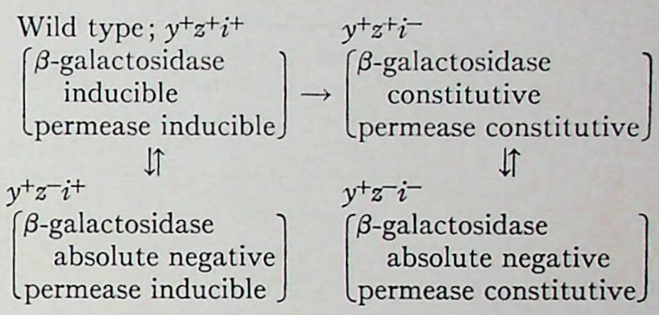
Before analyzing some of the hypotheses attempting to explain the mechanism of induction, it is useful to first describe the phenomenon of the constitutive synthesis of β-galactosidase in *E. coli* and the origin of the bacteria responsible for it.

If one examines a β-galactosidase constitutive strain of *E. coli* that is growing in the absence of an inducer, one will find that β-galactosidase synthesis occurs at a very high rate, 1000-fold or more greater than is found for the inducible strains growing under the same conditions. Furthermore, the addition of an inducer to such a constitutive culture does not augment this already high rate of synthesis. Thus, the constitutive strains synthesize β-galactosidase in the absence of inducer at rates which are apparently maximal for these cells and which are approximately the same as the maximum rates of induced β-galactosidase synthesis found in the inducible strains.

The question that is immediately raised in such a comparison of constitutive and induced β-galactosidase synthesis is whether one is concerned with the synthesis of the same protein molecule in each case or whether two different proteins are included in the above use of the term, β-galactosidase, one being synthesized in inducible strains and the other in constitutive strains. This question has been answered with some certainty in favor of identity.[3,19] Thus, the titration of catalytic units precipitable by a given quantity of specific anti-β-galactosidase sera yielded the same results with enzyme preparations derived from constitutive bacteria as with those derived from induced cells, whether the antibody was formed in response to the induced or to the constitutive enzyme. This not only demonstrates the antigenic identity of the two enzymes, but also indicates that their turnover numbers (catalytic units per molecule) are the same. Similarly, an extensive comparison of the kinetic constants for several substrates, for sodium- and potassium-ion activation, and for thermal inactivation of the β-galactosidase in these two preparations did not reveal any measurable difference.

Taking these observations in good grace as sufficient evidence for the identity of the β-galactosidase in constitutive and induced cells, one should expect the basic genetic unit or units necessary for the synthesis of this protein to be common to both cell types, assuming, of course, that one accept the tenets of the DNA-protein doctrine presented by Levinthal (p. 227). What then is the genetic difference between inducible and constitutive *E. coli*?

The following series of mutation sequences indicates the relationship between inducible and constitutive strains[16,20,21]:



In this scheme, the genetic units have the following meaning:

1. $i^+ \rightarrow i^-$ designates the mutation involving a change of state from inducible (i^+) to constitutive (i^-) for both permease and β-galactosidase.
2. $z^+ \rightarrow z^-$ designates the mutation involving damage to the z^+ unit necessary for β-galactosidase synthesis such that the enzyme cannot be made.
3. y^+ indicates the unit necessary for the synthesis of the galactoside-permease, a $y^+ \rightarrow y^-$ mutation (not shown) being that involved in formation of the cryptic, permease negative mutants previously mentioned.

This genotypic interpretation of the mutation data is consistent with data from recombination studies[22] and, as with a unique type of *cis-trans* test demonstrating that y^+ and z^+ involve different cistrons.[23]

From the fact that the β-galactosidase-proteins synthesized by constitutive and induced bacteria are identical, and from the genetic evidence demonstrating that the change resultant from mutation of inducibles (i^+) to constitutives (i^-) involves genetic units (cistrons) different from those involved in the formation of absolute negatives (z^-), one comes to the conclusion that the mechanism of synthesis of β-galactosidase in induced and constitutive cells does not differ in its essentials—that is, the same amino acids and the same mechanisms of ordering such amino acids are employed in each case. Though slightly more specific, this conclusion is equivalent to the general unitary hypothesis for enzyme synthesis emphasized by Cohn and Monod.[4]

Thus, in addition to conforming to the conclusions drawn directly from the induction phenomenon (Sec. II), any useful hypothesis for the induced enzyme synthesis must also explain the inducible to constitutive transition via the agency of a single-step mutation not involving a change in the basic genetic units determining the structure of the enzyme.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

IV. HYPOTHESES FOR INDUCED ENZYME SYNTHESIS

The point of departure for all hypotheses purporting to explain enzyme induction within the context of the unitary principle is the choice between two possibilities: (1) an inducer is an essential part of the minimum requirements for rapid enzyme synthesis (i.e., the "generalized induction principle"[4]); or (2) the converse, inducers are *not* part of these minimum requirements.

Since there exists no direct experimental evidence which determines this choice, it is necessary to analyze the consequences of both possibilities. A number of schemes have been developed that invoke the inducer as an essential element of the minimum requirements for enzyme synthesis.[4–6] It is not proposed to analyze these here, since they either do not attempt to explain at what point and in what manner the inducers act in the over-all conversion of amino acids to enzyme molecules,‡ or they are not consistent with the conclusions formed in Sec. II from data on β-galactosidase induction.§ Rather, a model of enzyme synthesis is presented here in which the necessity of inducer participation can easily be visualized and which is consistent with the conclusions drawn thus far. The initial assumption in this construction is that the inducer is neither necessary nor functional in determining the amino-acid order of the induced enzyme, but instead acts to increase the rate of formation of its secondary or tertiary structure. This assumption has its basis in the coding theory of DNA function in which the amino-acid order in a given protein is uniquely determined by the nucleotide order in the DNA of some functional gene (cf. Levinthal, p. 249).

A diagrammatic representation of how an inducer might exert a catalytic function in tertiary structure formation is given in Fig. 5. In this diagram, the semicircle of capital letters represents a template of unspecified chemical composition (DNA, RNA, or other) whose structure is determined by a functional gene and which in turn determines the order of amino acids (lower-case letters) making up a specific polypeptide. Each capital letter then represents a code symbol for one amino acid, and it is assumed that the physical structure which determines a given code symbol specifically binds only one amino acid. Peptide-bond formation between amino acids is imagined to occur on this template with the condition that the binding force between a given amino acid and its code structure is not lost by the formation of the polypeptide. Consequently, polypeptide and template form a relatively stable complex, which momentarily may assume any one of many partially dissociated states. Certain of these transitory, partially dissociated configurations will be favorable for the formation of the tertiary structure necessary for enzyme formation (represented here by the formation of a disulfide bond via the oxidation of two sulfhydryl residues in the polypeptide chain). The formation of such tertiary structure is assumed to be concomitant with dissociation of polypeptide and template, thus allowing the formation of free enzyme.

With no inducer present, such favorable states are considered to be extremely short-lived, so that the probability of formation of the disulfide bond pictured in Fig. 5 is quite small. As a consequence, the rate of enzyme synthesis is very low.

The assumed function of the inducer, or some product derived from it, (X), is to stabilize the favorable configurations by interacting with them to form a complex of the type represented by the middle left-hand structure in Fig. 5. Such a stabilization would result in an increase in the probability of the critical disulfide-bond formation and, therefore, in the rate of enzyme formation.

This mechanism is offered more as an aid in visualizing how one can imagine the necessity of inducer participation in protein synthesis, rather than as a unique representation of this possibility. However, it is consistent with the conclusions developed concerning β-galactosidase induction. Thus, it accounts for the *de novo* synthesis of β-galactosidase during induction, since

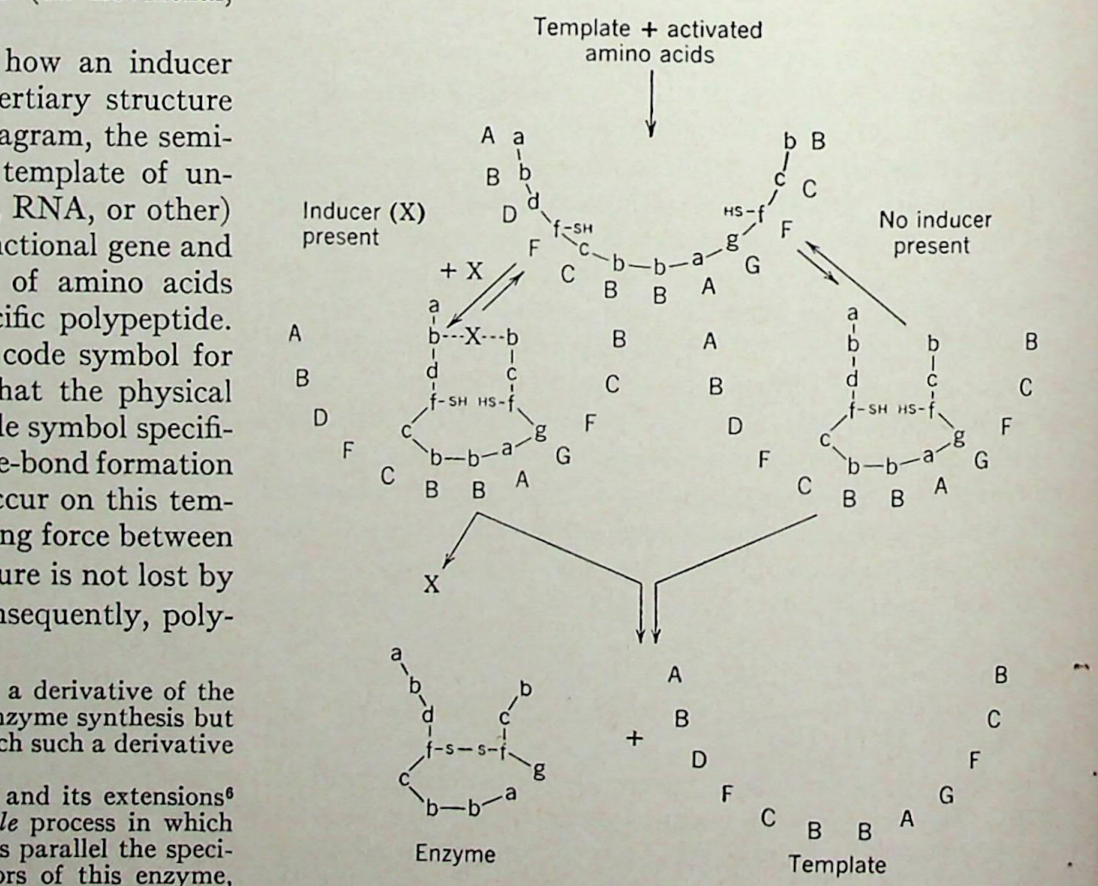


FIG. 5. Model for inducer function under the generalized induction hypothesis.

‡ The "organizer" hypotheses[4,5] insist that a derivative of the inducer (the "organizer") is an essential in enzyme synthesis but leave quite unspecified the mechanism by which such a derivative catalyzes enzyme synthesis.

§ The mass-action hypothesis by Yudkin[24] and its extensions[6] demand that enzyme induction be a *reversible* process in which the specificity of inducers of enzyme synthesis parallel the specificity of substrates and competitive inhibitors of this enzyme, conditions which are not consistent with the data derived from β-galactosidase induction (Sec. II).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

the level of polypeptide precursors of the enzyme in noninduced cells could be no greater than the number of templates per cell. Indeed, the calculated limit of ten or less such precursor molecules per noninduced cell is not an unreasonable limit for the number of templates per enzyme per cell. The model also allows for an essentially irreversible synthesis of β-galactosidase during the induction of growing populations; for the activation of enzyme-forming sites (templates) by inducers and the constancy of such sites during induction; and, finally, for the similarity but nonidentity of the specificity of inducers and substrates (or competitive inhibitors) of the enzyme. It is perhaps of interest to note that, insofar as reversal of the synthetic reaction in the model is limited by competition of activated amino acids and enzyme for the template (the enzyme having a very small affinity for the template as compared to the precursor polypeptide), the rate of any possible reverse reaction would be predicted to be greater under conditions of low levels of activated amino acids (e.g., in nitrogen-starved cells) than when these compounds are plentiful (e.g., in fast-growing cells.)

With respect to this model, or any model that invokes an inducer as an essential in the minimum requirements for the synthesis of a given enzyme, the two genetic states of inducible and constitutive β-galactosidase synthesis can be interpreted by assuming that constitutive cells contain an endogenous inducer, and that inducible cells are unable to synthesize such an inducer for lack of a necessary enzyme. On this basis, the mutation of an inducible strain to a constitutive can be viewed as a repair of a damaged genetic unit responsible for the synthesis of the postulated enzyme.

Turning now to the second possibility—namely, that an inducer is not a necessary component of the minimum system for enzyme synthesis—one must accept the constitutive condition as representing this minimum system. A model built for such a system could be like that given in Fig. 5, with two important differences: (1) The favorable state leading to tertiary structure could be stabilized by the form of the template itself, thus eliminating the need for inducer; and (2) the affinity of the activated amino acid for its code structure on the template would have to be lost as a result of peptide-bond formation with its nearest neighbors. The latter condition could be allowed for if one supposes that the activated amino acid consists of the amino acid covalently linked to a residue, *R*, which is unique for each amino acid, and that the code structure on the template has an affinity for *R* rather than the amino acid itself. If peptide-bond formation entails the breaking of the *R*-amino acid bond, then the polypeptide thus formed is quite free to separate from the template.

However, regardless of the model one builds for the synthesis of an enzyme without inducer (endogenous or exogenous) participation, the induced enzyme synthesis must differ from constitutive synthesis, either by a difference in type of template or by the existence of an inhibitor of normal synthesis (i.e., constitutive type synthesis) in the inducible strain, an inhibition which is relieved by the addition of an exogenous inducer. In the first of these possibilities, one must imagine that mutation from inducible to constitutive invokes a change in template type from one requiring an inducer for synthesis (i.e., as in Fig. 5) to one in which no inducer is required. Furthermore, since the β-galactosidase molecules synthesized constitutively and by induction are identical, one must suppose that this change in template type does not involve a change in the order of the amino-acid code. This seems to be asking a lot from a single mutation, but since almost nothing is known of the template structure, one cannot say, *a priori*, that such a change is impossible.

The second possibility—namely, that induction is the release of inhibition of enzyme synthesis—is actually quite an old idea, but in the past it was generally offered facetiously as the *bête noire* of generalized induction theories, since it could not be eliminated. However, recent developments demand more positive consideration of this idea. The source of these developments lies in the observation of what Vogel[25,26] has termed "enzyme repression." Repression is the inhibition of the differential rate of synthesis of an enzyme resulting from exposure of cells to a given substance ("repressor"). The word repression was adopted simply to avoid confusion with the term "enzyme inhibition," which by long usage implies the inhibition of the catalytic function of a given enzyme and not of its synthesis. One of the best examples of enzyme repression is observed among the constitutive enzymes included in the biosynthetic pathway leading to the synthesis of ornithine, citrulline, and finally of arginine in *E. coli*. Thus, Vogel[25] found that arginine is a specific and very effective repressor of the synthesis of acetylornithinase. Similarly, Gorini and Maas[27] have shown that arginine also represses the synthesis of ornithine transcarbamylase. Perhaps the most significant implication of these observations is that there exists within the cell a mechanism of feedback control between the result of the catalytic action of the group of enzymes in a biosynthetic pathway (in this case, the synthesis of arginine) and the synthesis of these enzymes. For present purposes, however, they give plausibility to the supposition that the inducible strains could represent a case of repression of the constitutive synthesis of β-galactosidase by endogenous repressors present in such strains, but not in the constitutive strains. Indeed, it is known that the constitutive synthesis of β-galactosidase in *E. coli* can be inhibited by exposure of cells to galactose or one of a variety of β-D-galactosides[4]; i.e., in the constitutive strains, these substances act as exogenous repressors of enzyme synthesis.

One may, therefore, term this second possibility the repressor hypothesis. It states that in constitutive strains enzyme synthesis occurs without the aid of any inducer; that such synthesis is inhibited in the inducible

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

cells because of repressor substances synthesized by these cells and not by constitutive cells; and that an inducer destroys this inhibition of enzyme synthesis either by combining with the repressor to yield a non-inhibiting complex, by competing with the repressor for some site on the template (or template-polypeptide complex)—which, if occupied by the inducer, does not inactivate the template, but if occupied by a repressor does so inactivate—or by inhibiting the synthesis of unstable repressors.

The three alternative explanations of induced and constitutive enzyme synthesis can be summarized as follows:

1. Generalized-induction hypothesis. Inducers are necessary components for enzyme synthesis both in constitutive and in inducible cells. Endogenous inducers are synthesized by constitutive cells but not by inducible cells, which must, therefore, receive exogenous inducers for enzyme synthesis.

2. Different-template hypothesis. Two different types of template exist for enzyme synthesis. In constitutive cells, the templates function without the aid of inducers. In inducible cells, the inducers are necessary to activate the template (as in Fig. 5). No endogenous inducer or repressor is assumed. It is assumed that the amino-acid ordering function of each template is the same.

3. Repressor hypothesis. The templates in inducible and constitutive cells are the same and do not require activation by inducers. Substances specifically inhibiting template function (repressors) are synthesized by inducible cells but not by constitutive cells. Exogenous inducers function by destroying this inhibition of enzyme synthesis.

V. EVALUATION OF THE INDUCTION HYPOTHESES

The data presented thus far do not allow much more than an intuitive preference for one of the three hypotheses over the others. Thus, the enzyme repression observed in the enzymes involved in arginine biosynthesis can equally well be explained by the inhibition of endogenous induction as by direct inhibition of template catalysis not involving an inducer. However, experiments recently reported by Pardee *et al.*[23] have done much toward clarifying the relative weight that one can place on the three alternative explanations of enzyme induction. These experiments were designed to observe the behavior of β-galactosidase synthesis in *E. coli* zygotes having a heterogenotic structure for both the z^+, z^- and the i^+, i^- pairs in an effort to determine dominance in the i^+, i^- pair. This determination is important in the evaluation of the hypotheses, since, in the generalized induction hypothesis as it has been presented, the positive and, therefore, dominant function of synthesizing an endogenous inducer is given only to the constitutive (i^-) strain, whereas in the repressor hypothesis the positive and, therefore, dominant function of synthesizing a repressor is given only to the inducible type (i^+). In the hypothesis of two template types, both inducible (i^+) and constitutive (i^-) units are given positive functions and, therefore, no dominance in the i^+, i^- pair should necessarily exist.

The characteristics of *E. coli* conjugation allow the determination of dominance in a very unique manner. The male (*Hfr*) member of the mating pair injects its "chromosome" into the female (F^-) member through a small tubule connecting the two cells. The order in time of entrance into the female of the genetic units from a given strain of male cells is unique, and apparently no appreciable amount of cytoplasm enters with these genetic units (cf. Wollman *et al.*[28] and the chapter by Lennox p. 242). Thus, the initial cytoplasmic state of the zygote is determined by the cytoplasmic state of the female before conjugation. As a consequence, it is possible to form the i^+z^+/i^-z^- zygote either by the injection of the i^+z^+ genetic units from an inducible male into an i^-z^- formed cytoplasm of a constitutive absolute negative female or vice versa—to inject the i^-z^- genetic units into an i^+z^+ formed cytoplasm. The question then asked is what is the behavior of β-galactosidase synthesis in the i^+z^+/i^-z^- zygote in each case and to what hypothesis does this behavior correspond? In answer to this question, it is convenient to first describe the behavior predicted by each hypothesis.

1. Under the form of the generalized-induction hypothesis presented here, the mating of a male i^+z^+ cell to a i^-z^- female in the absence of any exogenous inducer should yield the synthesis of β-galactosidase soon afte the i^+z^+ units have been injected into the i^-z^- cytor plasm, since this cytoplasm should contain endogenous inducers which can activate the template resulting from the entering z^+ unit. It should be noted here that the z and i units are very closely linked[23,29] so that, within the minimum time units employable in conjugation experiments, the two units enter simultaneously. Furthermore, the synthesis of β-galactosidase should continue in the i^+z^+/i^-z^- zygote as long as this heterogenotic state remains intact.

In the reciprocal conjugation, the prediction would be the same, except that there might be expected to be a longer lag between the entrance of the i^-z^- units into the i^+z^+ cytoplasm and the first appearance of enzyme synthesis owing to the necessity of the entering i^- unit to catalyze the synthesis of the endogenous inducer in the zygote cytoplasm.

2. Under the hypothesis involving two types of templates for constitutive (i^-) and inducible (i^+) synthesis, the i^+z^+/i^-z^- zygote should yield no β-galactosidase synthesis in the absence of exogenous inducer, unless the z^+ in one "chromosome" and the i^- in the other can cooperate to yield the type of template synthesized in the constitutive β-galactosidase positive strain (i^-z^+). If this is possible, then the β-galactosidase synthesis should occur in the zygote with the same behavior, whether the i^+z^+ units are injected into a i^-z^- female or

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

vice versa. It should be noted that the possibility of recombination leading to i^-z^+ on the same "chromosome" is negligible owing to the closeness of the z and i units.

3. Under the repressor hypothesis for induction, the injection of a i^+z^+ unit into a i^-z^- cytoplasm in the absence of inducer should yield the synthesis of β-galactosidase soon after the injection, since initially there should be no repressor in the zygote cytoplasm, it being of the constitutive (i^-) type. However, different from the prediction of the generalized induction hypothesis, in this case one should expect the synthesis of β-galactosidase in the zygote to cease as soon as the injected i^+ unit has caused the synthesis of sufficient repressor in the zygote cytoplasm. At this time, the zygote should be phenotypically inducible rather than constitutive as it was immediately after the entrance of the i^+z^+ unit.

In the reciprocal mating, the cytoplasm of the zygote, being i^+z^+, should initially contain the repressor and should continue to synthesize repressor after the i^+z^+/i^-z^- state is established. Thus, no synthesis of β-galactosidase in absence of exogenous inducer would be expected at any time after conjugation.

Each hypothesis yields a different prediction, and, consequently, the experiment should uniquely determine which, if any, is valid. The results are entirely consistent with the repressor hypothesis. Thus, in the conjugation of i^+z^+ males with i^-z^- females in the absence of exogenous inducer, β-galactosidase synthesis can be detected within a few minutes after the entrance of the i^+z^+ genetic units. However, in the reciprocal conjugation, no enzyme synthesis could de detected in the absence of inducer, even after several hours. Furthermore, in the i^+z^+ male-to-i^-z^- female conjugation the synthesis of β-galactosidase in absence of inducer ceases about two hours after zygote formation. This is shown in Fig. 6. This cessation of synthesis is not the result of segregation of the zygote, since this event cannot be detected until two hours after cessation of synthesis. As is shown in Fig. 6, the zygotes have become inducible by the time the constitutive synthesis stops. These results are exactly those predicted under the repressor hypothesis in which i^+ was predicted to be dominant over i^-, and they are inconsistent with the two other hypotheses. These results, furthermore, offer the evidence promised earlier that the i and z units involve different cistrons—i.e., different functional genetic units.

Does this experiment then offer the death knell to the hypothesis of different templates and to the generalized induction theory? It quite effectively eliminates the former of these possibilities, but, unfortunately, the generalized induction hypothesis can be made viable again by a very slight alteration. The supposition made under this hypothesis was that the i^+ to i^- mutation involved the repair of a genetic unit necessary for the synthesis of endogenous inducer. This supposition is clearly eliminated by the experiments of Pardee, Jacob, and Monod. However, it can be assumed that, in both

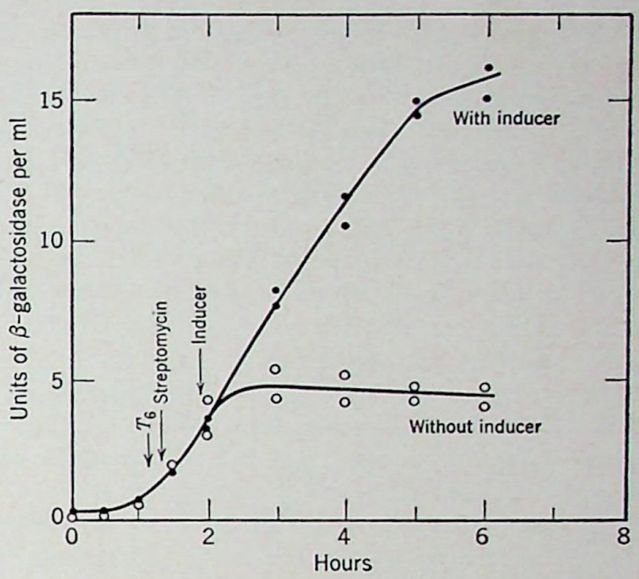


FIG. 6. Formation of β-galactosidase in zygotes from a $\male i^+z^+ \times \female i^-z^-$ conjugation of *E. coli*. The abscissa indicates the time after mixing the parental populations. The experiment was performed in quadruplicate. Bacteriophage T_6 and streptomycin were added at the indicated times to stop conjugation and to eliminate the ♂. The inducer, methyl-β-D-thiogalactoside ($10^{-3}M$), was added at 115 min to two of the four cultures (filled circles), whereas the other two cultures (open circles) contained no inducer [from A. B. Pardee, F. Jacob, and J. Monod, Compt. rend. **246**, 3125 (1958)].

i^+ and i^- cells, an endogenous inducer is synthesized, but that i^+ cells differ from i^- cells in possessing an enzyme capable of destruction of this endogenous inducer. The i^+-to-i^- mutation would then result in the loss of ability to synthesize this enzyme, and i^+ would be expected to be dominant to i^-. The results of the above experiment are equally well in accord with this extension of the generalized induction hypothesis as with the repressor hypothesis. Thus, this ingenious experiment has indeed limited the choice of hypotheses drastically; it has pushed the repressor hypothesis fully into the limelight, but it unfortunately does not offer a final unambiguous solution.

Nor should one expect such an unambiguous solution as long as one is forced to deal with the complexity of an entire cell. At this level, a solution will be unique only in that it unifies and poses the answer to a larger number of problems with a lesser number of assumptions than does its alternatives. In this sense, the repressor hypothesis is the more satisfying. It yields a mechanism for feedback control of enzyme synthesis that would seem to be necessary if the cell is not to run amuck. If the phenomenon of enzyme repression becomes a general observation, particularly in biosynthetic pathways such as that found in arginine synthesis, one can account for feedback control with the use of one or less repressor substances per enzyme. However, in choosing the generalized induction hypothesis, if one wishes to maintain the explanation of feedback control, one must involve inducers and substances inhibiting induction (i.e., repressors) for each enzyme. Thus, if both hypotheses are generalized, it would seem that

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

REVIEWS OF MODERN PHYSICS VOLUME 31, NUMBER 2 APRIL, 1959

32

General Features of Radiobiological Actions*

RAYMOND E. ZIRKLE

Committee on Biophysics, The University of Chicago, Chicago 37, Illinois

THIS paper and the three that follow (Pollard, p. 273; Wood, p. 282; Tobias, p. 289) are concerned with radiation biology. The potential scope of this field—effects of all types of radiations on all types of biological systems—and its relation to other areas of radiation research are indicated in Fig. 1, where an energy spectrum of radiations is plotted as abscissae and various inanimate and biological systems are "plotted" as ordinates in ascending order of presumed complexity. At the level of macromolecules, radiation biology shades imperceptibly into radiation chemistry and physics, the basic sciences on which it draws for necessary facts and concepts. At the level of groups of multicellular organisms, it approaches "radiation sociology," which is needed to deal with the problems of group behavior in irradiated populations.

Radiations below about 1 ev produce biological effects, if any, through heating. These effects usually are regarded as cognate to, but not included in, radiation biology. The region from 1 to 6 ev, comprising the near infrared, the visible, and the familiar ultraviolet wavelengths, is readily available for investigations as is the region from about 1000 ev upward. These two accessible portions of the spectrum conveniently are termed low-energy and high-energy, respectively. The intervening "transition" region is absorbed in air and biological material so readily that it can be used only on very small objects *in vacuo*. It is of great theoretical interest, in view of the great differences in basic response of biological systems to the low- and high-energy spectra.

Figure 1 shows that radiation biology potentially ramifies throughout biology as a whole. So does biophysics. What is the relationship, if any, between the two? In my personal view, each contains some of the other. The biophysical content of radiation biology is probably about as great, both potentially and actually, as that of general biology.

It is clearly impracticable, in four short papers, to cover all of radiation biology. Accordingly, the coverage is narrowed as follows. First, attention is concentrated on the high-energy radiations, with occasional comparative references to the ultraviolet. Second, specific samples are selected from the spectrum of biological systems (Fig. 1); after these brief general remarks, Pollard describes experiments on certain macromolecules and viruses (p. 273), Wood speaks on certain cellular effects that have been investigated intensively (p. 282), and Tobias discusses selected studies on cell populations and multicellular organisms (p. 289).

By and large, the effects of high-energy radiations, as well as those of the ultraviolet, are injurious to all or part of the irradiated system. Nevertheless, considerable effort goes into the study of these effects, for various reasons. Some persons find a fascinating field of research in the general problem of how very small amounts of radiation energy produce such drastic effects. I believe that the real attraction here is the peculiar combination of various portions of physics and chemistry and of many aspects of biology that must be brought to bear on any serious investigation of radiobiological mechanism. This peculiar versatility demands so much of a human life span that a radiation biologist is necessarily a specialist, much as he may strive to not become too differentiated.

However, most of the interest in radiation effects stems from their applications. In many branches of basic biology, they have long been used as powerful research tools. For example, the development of genetics has been accelerated immensely by the use of radiations as mutagenic agents; partial body irradiation has found fruitful application in embryology; partial cell irradiation is used to get information about the properties and functions of various cell parts. As basic biological tools, the various radiations have two properties that, in many situations, are of critical advantage: they do not disrupt membranes and other structures grossly, and the dosage usually is reproducible.

Radiation biology also has some important practical applications. The oldest, and still one of the most important, is radiation therapy. Another is its use in dealing with the widespread and multifarious problem of radiation hazards that, within the last decade and a half, have increased so explosively that they even figure in national and international politics.

Regardless of one's motivation to study or use radiation effects, there is obvious need to know as much as possible about their mechanisms. There is nothing basically unique about current methods of investigating these mechanisms; they are essentially those of the physical sciences and of the analytical biological sciences, with some special variations that stem from the unique physical properties of the high-energy radiations.

Like any story, a radiation action has a beginning, an end, and a middle. The beginning is the act of irradiation, and the end is the effect observed; there is considerable information about these, and the prospects of

* The following data are used in this paper: The usual energy unit is the electron volt (ev); 1 ev $=1.6\times10^{-12}$ erg $=3.8\times10^{-20}$ cal. For any radiation, energy times wavelength is 1.235×10^{4} ev A. "Dose" is radiation energy transferred to unit mass of irradiated object; 1 rad $=100$ ergs/gram $=66$ ev/μ^3, assuming density of object to be 1.05; 1 roentgen (r) *usually* is equivalent to 0.93 rad.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

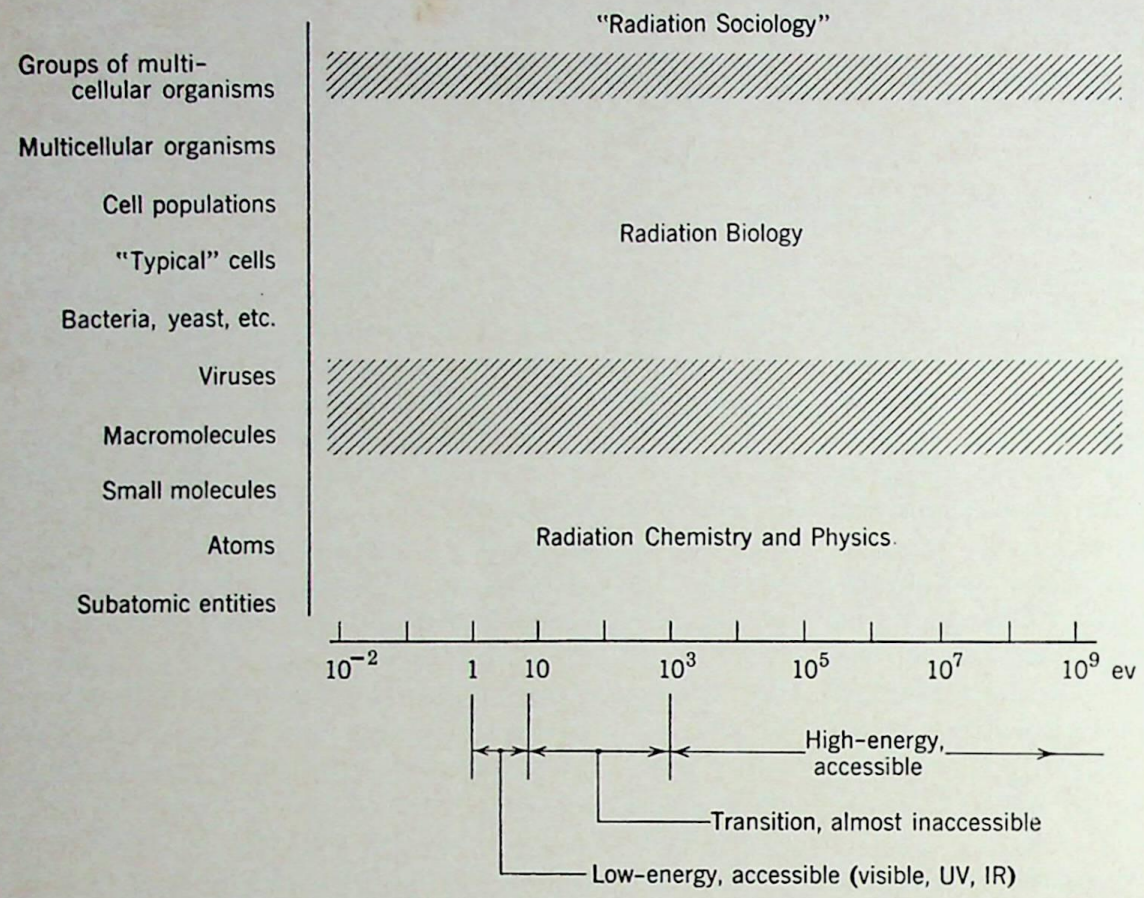


FIG. 1. Scope of radiation biology and its relations to other areas of radiation research. Abscissae, partial radiation energy spectrum in electron volts; ordinates, "spectrum" of inanimate and biological objects in order of presumed complexity.

getting more are good. The middle, frequently miscalled the "latent period," is essentially a domain of ignorance wherein most of the problems lie. The remainder of this article is devoted to an attempt to indicate some of the ways in which current research is directed to reduction of this ignorance.

Clearly, the end effect must be defined as accurately as possible. This implies information about the state of the investigated system *before* it is irradiated. The incompleteness of such information is demonstrated well by the need for a large conference such as the Study Program in Biophysical Science and by the remedial attempts so vividly described by earlier speakers. High-energy radiation in sufficient amount is capable of producing significant changes in practically any biological structure or function. In view of the great diversity and complexity of these structures and functions, it is not surprising that the list of known end effects is long. Some vigorously investigated examples are aberrations in chromosome structure, inhibition of cell division, inactivation of enzymes and other macromolecules, gene mutations, and induction of tumors in animals.

It is also essential to know as much as possible about the process of irradiation and the immediately consequent events. Here, one encounters facts and concepts that are not basic to other areas of biological research. The high-energy radiations differ in two important features from all other physical and chemical agents, including the low-energy radiations. First, a large fraction of the molecules that receive energy from the radiation are activated to very high states, even to ionization. As a result, in most biological material, radiation energy is transferred to the individual molecules of the many species usually present in a nonselective fashion that is chiefly dependent only on the molecular mass. Chromophores do not enter into consideration, as they do in work with low-energy radiations. This circumstance presents both advantages and disadvantages to the investigator.

Second, the energy transfers to individual molecules do not occur singly in space and time but in more or less linear groups. All of the high-energy radiations are either charged particles in motion (e.g., electrons or α-particles) or are agents (neutrons; x-rays and γ-rays) that set such particles in motion. These charged particles typically have kinetic energies of thousands or millions of electron volts, whereas the average energy that can be accepted by an individual molecule is some tens of electron volts. Thus, successive energy transfers occur along the path of the high-energy particle. Some of these transfers are capable merely of exciting the molecules to higher energy states. Others involve enough energy to produce ionization, e.g., to eject electrons, some of which have kinetic energies sufficient to eject electrons from other molecules (secondary ionization). The ions produced in gases are well demonstrated by means of the Wilson cloud chamber. The trail of ions produced by a single primary charged particle is termed an ionization track. Similar tracks are recorded on photographic plates (linear sequences of blackened grains) and in suitable liquids at low temperatures (trails of bubbles). This, and other physical evidence, makes it highly probable that the distribution of energy transfer in a biological object is essentially the same as that dis-

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

played in the cloud chamber, except that the dimensions should be reduced by a factor of about eight hundred because of the difference in density.

These two properties—the high average value of the individual energy transfers and the grouping of these transfers into tracks—make the high-energy radiations unique in their mechanisms of action, even though these mechanisms, in some cases, produce end effects that are superficially indistinguishable from those of other agents.

To students of mechanism, probably the most significant single feature of high-energy radiation actions, especially on cells and on even more-highly organized systems, is the small amount of energy required. Of the many end effects that are known, most are produced to a significant degree by 10 000 rad (or less). This dose equals about 0.02 cal/g of biological material.

For some actions, the dose-effect relations ("kinetics") are suggestive. When samples of certain small cells (bacteria, haploid yeast, etc.) are given graded doses, the resulting survival curves are exponential; for more-complicated cells and organisms, the curves are sigmoid. The exponential curves suggest that the action on each cell is owing to some single primary event (formation of an ion pair, passage of a single ionizing particle, etc.). Sigmoid curves are correspondingly ascribed to multi-event types of action. In some cases, as Wood points out, the curve is exponential or sigmoid, depending on known properties of the cells. This is useful in devising models for the action.

Macromolecules and viruses typically exhibit exponential curves, although their shapes may be modified by various factors. By making simple assumptions concerning the nature of the hypothetical single event, the size of the "target" relevant to the radiation action can be calculated[1,2]; Pollard describes this procedure in detail (p. 273).

In dilute inorganic aqueous solutions, the alteration of the solute (e.g., oxidation of ferrous ion) typically follows zero-order kinetics, which indicates that the energy is absorbed principally by the solvent and that chemical intermediates, formed from solvent molecules, react with the solute. If two or more solutes are present, they compete for the intermediates and thus "protect" each other. Actions involving intermediates are termed "indirect."[3] "Direct" actions are those in which the radiation energy must be transferred to the solute molecules themselves.[4]† Current radiation chemistry identifies some of the intermediates[4,5]; they are OH and H radicals, H_2O_2 molecule, and, in the presence of molecular oxygen, HO_2 radical.

All of these concepts, derived from work on simple solutions, ramify throughout radiation biology. Indirect action has been demonstrated in many radiation effects on macromolecules in solution.[3] If a preponderant concentration of protective substances is present, a residual effect is observed that is ascribed to direct action. Indirect action has been invoked in analysis of various cellular effects[6]; the concept has been extremely fruitful in suggesting experiments, although in no case has its correctness been established for systems as complex as cells.

Many physical, chemical, and biological factors are known to modify the amount of absorbed radiation energy necessary to produce a given degree of effect. A few that have been found to operate in a wide variety of systems are mentioned here.

One is the physical parameter called linear energy transfer (LET), which is the amount of energy transferred per unit length of track of an ionizing particle to the molecules it traverses. This parameter varies with the square of the charge on the particle and increases as the velocity decreases. With the various kinds of radiation currently available, one can obtain LET values ranging from 0.025 to 25 ev/A, corresponding roughly to average spacings between primary ionizations that range from 4000 down to 4 A. Thus, it is possible to give the same dose to a biological system in various ways: a few ionizing particles with high LET, many particles with low LET, and intermediate values. Results are detailed elsewhere.[7] Not only does LET significantly influence the dose-effect relations in practically all radiobiological actions that have been investigated thoroughly in this respect, but it also operates in "simple" chemical actions, such as the formation of H_2O_2 in pure water. Thus, LET not only gives some geometrical notions about mechanism, but also gives some encouragement to use radiation chemistry as a basis for interpretation of radiobiological actions.

Another factor quantitatively influencing a wide variety of actions is molecular oxygen.[8] In all but a few of these actions, radiosensitivity increases with concentration of O_2 until a plateau is reached, usually at a value two or three times that observed when the O_2 concentration is zero. The basis of the "oxygen effect" is still controversial (cf. Wood, p. 282), but I think it significant that, like LET, it is encountered in radiation effects on simple aqueous solutions. There is also an interrelationship between O_2 and LET: the greater the LET produced by the radiation, the less is the influence of O_2.

In the foregoing, I have tried to communicate a concept of the present scientific state of basic high-energy radiation biology. Many effects on many diverse biological systems are known and cataloged; however, there is encouraging evidence that the mechanisms leading to these diverse effects have some strong resemblances. No one mechanism has been elucidated yet. On the other hand, several have been investigated intensively by means of the general approaches indicated, and, in a few cases, observed facts have been used as bases for theories which have been successful; i.e., they

† If no solvent is present, the action must be direct. Target theory, in its strict sense, presupposes direct action.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

have suggested experiments which in turn have yielded new, significant facts. Good examples of such investigations are given in the three papers which follow.

BIBLIOGRAPHY

[1] D. E. Lea, *Actions of Radiations on Living Cells* (Cambridge University Press, London, 1955), revised edition.

[2] E. C. Pollard, W. R. Guild, F. Hutchinson, and R. B. Setlow, Progr. in Biophys. and Biophys. Chem. **5**, 72 (1955).

[3] W. M. Dale in *Radiation Biology*, A. Hollaender, editor (McGraw-Hill Book Company, Inc., New York, 1954), Vol. I, p. 255.

[4] A. O. Allen, J. Phys. & Colloid Chem. **52**, 479 (1948).

[5] J. L. Magee, Ann. Rev. Nuclear Sci. **3**, 171 (1953).

[6] R. E. Zirkle and C. A. Tobias, Arch. Biochem. Biophys. **47**, 282 (1953).

[7] R. E. Zirkle in *Radiation Biology*, A. Hollaender, editor (McGraw-Hill Book Company, Inc., New York, 1954), Vol. I, p. 315.

[8] Z. M. Bacq and P. Alexander, *Fundamentals of Radiobiology* (Academic Press, Inc., New York, 1955).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

REVIEWS OF MODERN PHYSICS VOLUME 31, NUMBER 2 APRIL, 1959

33

Radiation Inactivation of Enzymes, Nucleic Acids, and Phage Particles

ERNEST POLLARD

Department of Biophysics, Yale University, New Haven, Connecticut

INTRODUCTION

IT is worthwhile, before the description of experimental work, to say a word about the motivation of the researches to be described and also something of the people engaged in them. The purpose has been to use ionizing radiation as a powerful, localized, and penetrating agent to study cell structure in relation to cellular function. This purpose was clearly in the mind of the late D. E. Lea, and in many ways we have been continuing lines which he began. The purpose can be usefully directed only if there is some knowledge of the actual character of ionizing radiation action on the three key elements of cellular systems: proteins, nucleic acids, and polysaccharides. Such knowledge is still imperfect, and what is here described is a series of studies, which enable one to make some preliminary hypotheses as to the action of ionizing radiation. The promise of immediate future progress is excellent, and, if research effort on the right scale were forthcoming, a year or two more would see the major features of the pattern properly exposed.

The people in the Yale Biophysics Department who have contributed are Hutchinson, Setlow, Guild, Preiss, Woese, Powell, Forro, Fluke, Jagger, Wilson, Till, and Whitmore. The author writes as representing this group and the work is theirs as much as his in every respect.

In the process of studying the inactivation of these key molecules, the basic problems of radiobiology are also being studied; and since radiobiology cannot be separated from biology, the problems of biology are also involved. This will become apparent, as has been stressed already by Zirkle (p. 269).

GENERAL CHARACTER OF RADIATION INACTIVATION

The physical action of radiation is complex. All optically allowed molecular transitions presumably are capable of occurring, together with many forbidden transitions. The gradual dispersal of the intense local energy releases, and the accompanying "pre-equipartition" local high-temperature regions, plus the probable effect of local high temperatures, conspire to give the molecular physicist a choice of almost any personally favored mechanism of action. Such choices have been made, and it would be foolish to exempt this paper. In view of this great complexity, the traditional physical introduction is foregone and the experimental results on biological macromolecules are discussed directly.

The sort of experiment which is readily performed requires a set of assay tubes for some enzymes and a color agent that is dark for active enzymes. Since the reaction time is kept constant, the activity of the enzyme shows in the relative color density, and, for an enzyme heavily irradiated with γ-rays, the activity is clearly low. Such experiments were carried out first by Northrop[1] and by Hussey and Thompson.[2] Similar experiments on viruses are described later. Interestingly enough, the earliest recorded quantitative measurements on bacteriophage were made by the Sinclair Lewis hero, Martin Arrowsmith, which probably reflects the active mind of Paul de Kruif.

The dose-response curve found in these experiments generally obeys the relation

$$\ln(n/n_0) = \text{const} \times \text{dose}. \tag{1}$$

It can be explained statistically very simply, as suggested by Dessauer,[3] Crowther,[4] and Condon and Terrill[5] many years ago, by supposing that I inactivating events per unit volume are distributed randomly and that there is a critical sensitive volume V which may intercept one of these events. Since the average number of events in a volume V is IV, then by the Poisson formula the probabilities of 0, 1, 2, 3, etc., events taking place in the volume are $P(0)$, $P(1)$, $P(2)$, $P(3)$, where

$$P(0) = e^{-IV} \qquad P(m) = e^{-IV}(IV)^m/m!$$

$$P(1) = e^{-IV}IV$$

$$P(2) = \frac{e^{-IV}(IV)^2}{2!}$$

These are mutually exclusive events so that $\sum P(m) = 1$, and, accordingly, one can reason thus: If the probability of escape is measured by the ratio of the number left active, n, to the number at the start, n_0, then for complete escape, $P(0)$, one has

$$n/n_0 = e^{-IV}. \tag{2}$$

If one "hit" can be withstood,

$$n/n_0 = e^{-IV} + e^{-IV}IV. \tag{3}$$

If two hits can be withstood,

$$n/n_0 = e^{-IV} + e^{-IV}IV + \frac{e^{-IV}(IV)^2}{2!}, \tag{4}$$

and so on.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

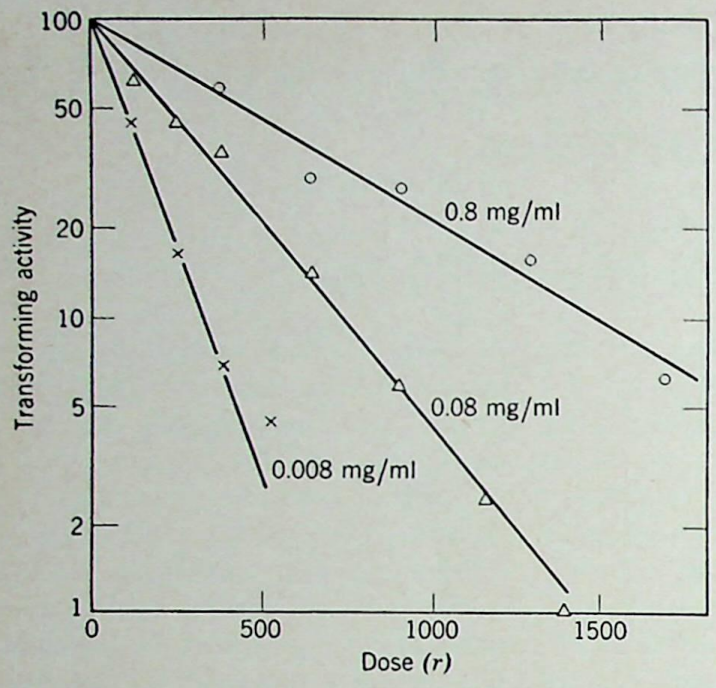


FIG. 1. The loss of transforming activity in pneumococcus DNA preparations in dilute solution. This is typical of irradiation in dilute solution where the majority of the effect is owing to the migration of radicals. That this is so can be seen from the increased inactivation in more dilute solution. Data due to DeFilippes and Guild.[6]

Usually, for enzymes and viruses, one finds the simplest, complete escape expression holding, or

$$\ln(n/n_0) = -IV. \tag{5}$$

Thus, I can be identified somehow with the dose and V with the constant in Eq. (1). The data of Figs. 1 and 2 provide examples of two inactivations which obey this relation. Figure 1, taken from work by DeFilippes and Guild,[6] shows the effect of x-rays on the transforming principle of pneumococcus, which Hotchkiss has shown to be pure DNA, and which was referred to by Rich (p. 191). The ordinate, the activity, is plotted logarithmically *vs* the dose in roentgens, and the relation for a one-hit effect holds. If the inactivation is ascribed to primary ionizations per unit volume, using a conversion factor of $5\times10''$ primary ionizations per cubic centimeter in water, an absurd value for V is found, moreover one which depends upon the concentration of transforming principle (TP). The inactivation is due clearly to "activated water," probably free radicals which can migrate and so make V many times larger than the volume of the molecule. In principle, if the radicals in activated water have infinite lifetime, and if the water and TP are pure, there always should be one molecule of TP inactivated per radical. In fact, radicals recombine; Hutchinson and Ross[7] and Smith[8] have estimated their half-life. In very pure water, it is as long as 10^{-4} sec. In yeast, it is 10^{-9} sec.

As a contrast, Fig. 2 shows some new data, taken by the author and Nancy Barrett, for inactivation of β-galactosidase. With one exception, the irradiations were in the dry state. All were done with a cobalt γ-ray source. The black dots show the effect of radiation on enzyme extracted from lactose-adapted bacteria. The relation $\ln(n/n_0) = -IV$ is obeyed. If one uses for I the number of primary ionizations, or clusters of ions, one finds $V = 4.7\times10^{-19}$cc. Since one usually is not familiar with the volumes of molecules, this figure can be converted into a molecular weight by assuming a protein density of 1.3 and multiplying the mass of one molecule by Avogadro's number to get the molecular weight. The figure found is 370 000. Estimates for the molecular weight of β-galactosidase are mostly guesses, but, if one takes the molecule to be spherical, a sedimentation constant of 18, which has been quoted, and which is approximately checked by work in our laboratory by Langridge, gives a molecular weight of 390 000. So, in this case, the dry measurements lead to a statistical

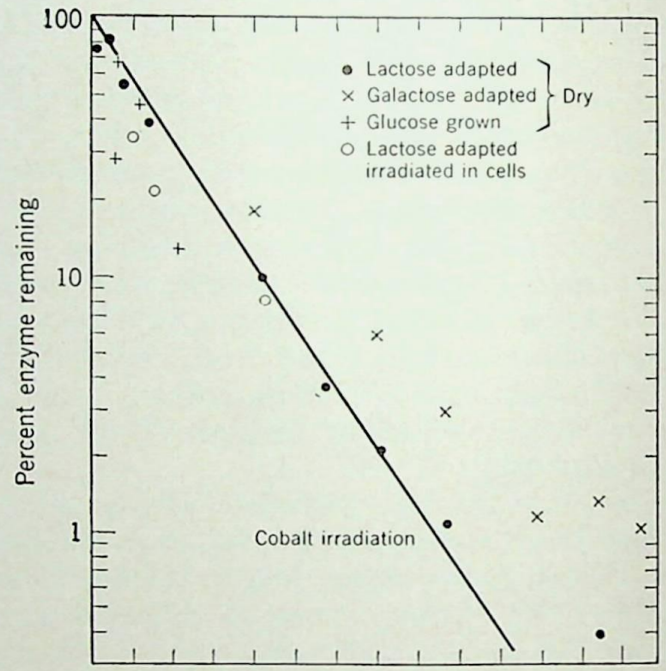


FIG. 2. The irradiation in the dry state of β-galactosidase grown under various conditions. The characteristic behavior is that an essentially uniform method of inactivation is observed dependent only on the direct hitting of some sensitive region. Irradiation in the wet state is put in for comparison. It is clear that in the wet state inside the bacterium the concentration is so high that inactivation is similar to the dry inactivation.

analysis in which the volume V may well be the volume of the molecule.

The "constitutive" enzyme, that small fraction which is present in unadapted cells, behaves about the same as does enzyme from galactose-adapted cells. The sensitivity of the enzyme in cells irradiated in C-minimal medium is not noticeably different.

Thus, one may draw radiation action on a molecule in two ways, as shown in Fig. 3. The shaded part in the center is possibly identifiable with the macromolecule itself, while the dotted lines are owing to migration of active chemical agents. For the purpose of the study of biological systems, the system can be forced very often into conditions where only the shaded part is operative. Sometimes this cannot be done, a familiar restriction to biologists.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

QUANTITATIVE STUDIES ON PROTEINS

The result that the "inactivation volume" V is closely related to the volume of the molecule first was suggested in work by Lea *et al.* in 1944.[9] In our laboratory, their work has been extended greatly to include many enzymes, albumin, and hormones; contributions by others, notably Fluke,[10] also have been made. A graph summarizing these, owing to Guild, is shown as Fig. 4. A log plot of the "radiation molecular weight" *vs* the accepted molecular weight is shown. There are some clear deviations, but the relation hardly can be overlooked. It is most remarkable and unexpected, and for some time no reasonable explanation could be given. In 1955, the suggestion was put forward[11] that, in a covalently bonded structure, there would be no reason why an initially positive region produced by ionization should be confined to one atom, but rather it should migrate and settle possibly in some point of weakness. Recent work by Gordy[12] using paramagnetic-resonance hyperfine structure has shown that, in irradiated material as extensive as proteins, there can be excited two characteristic patterns, one of which is clearly identified as owing to cysteine. Figure 5 shows some of his results.

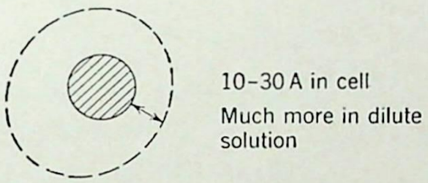


FIG. 3. Indication of the distance of migration of radicals in a cell. It has been estimated by Hutchinson that an increase in distance of 10 to 30 A corresponds to the distance of migration of a radical through a sensitive target. In dilute solution the distance of migration is much greater.

That irradiation, which hardly can have been so intense as to excite the whole set of amino acids, should consistently give such patterns, strongly suggests that migration of the positive charge indeed does occur, resulting in its statistically settling in one or two favored spots. We propose to call this the *primary lesion*.

Subsequently, the protein is exposed to water or to water and oxygen, and the primary lesion now exhibits a reactivity which is called the *chemical action*. This chemical action, by breaking an $-S-S$ bond, or by removing a side chain, gives an altered molecule. We thus feel that, as of 1958, the process is to be regarded as in Fig. 6. On this view, there should be a possibility of modifying the action at two steps, A and B, and a lesser possibility at a. It has been shown that invertase and catalase are markedly more sensitive at temperatures just below the thermal inactivation region. This is possibly action at a or A. Braams, Hutchinson, and Ray[13] have shown that ribonuclease dried in acetic acid is four times as sensitive, and in glucose, twice as sensitive as normally found. A variety of additives have no effect, among them salt and glycine, while glutathione

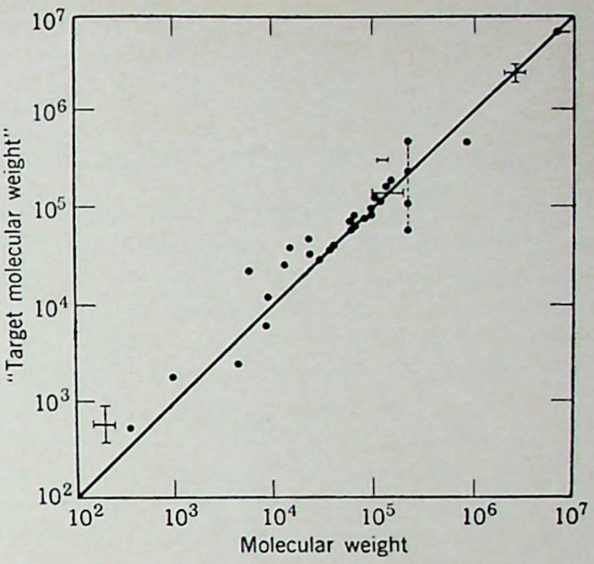


FIG. 4. A plot comparing the observed molecular weight with the target molecular weight, for a wide variety of bombarded substances. This graph has been prepared by Guild and shows the plausibility of the idea that a primary ionization anywhere within the molecule will cause the loss of biological activity. There are exceptions and quite clearly there can be means of influence which will alter the sensitivity of the target, but by and large there is a good relationship.

reduces the sensitivity by a factor of two. A sensitivity modification which is eightfold thus can be achieved. Hutchinson[14] also has confirmed the finding of Alexander[15] that dry trypsin irradiated in air is 2.3 times as sensitive as in nitrogen, for γ-radiation. For deuteron or α-particle irradiation, the effect is far less.

The fact that the relation shown in the log plot of Fig. 4 holds, means that most additives which occur in ordinary separation procedures are fairly uniform and not able to produce the effects described above.

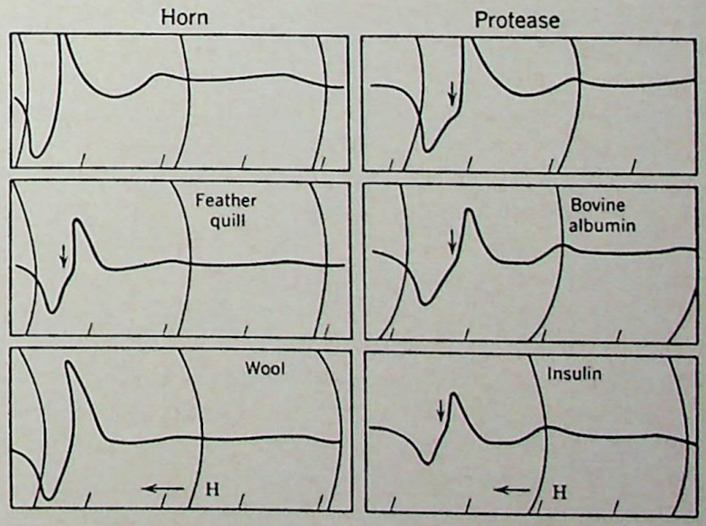


FIG. 5. Reproduction of data due to Gordy showing the paramagnetic resonance in a number of irradiated proteins. It is significant that while a variety of substances are shown quite similar resonance is observed in many of them. This means that there are probably regions in the molecule at which the formation of an unpaired electron is preferred and these can be regions of weakness for chemical action later on [from W. Gordy, W. B. Ard, and H. Shields, Proc. Natl. Acad. Sci. U. S. 41, 983 (1955)].

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

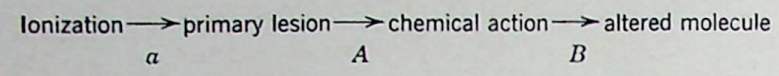


FIG. 6.

QUANTITATIVE STUDIES ON NUCLEIC ACIDS

The transforming principle is pure DNA, and so affords a possible (though not very convenient) measure of the biological activity of DNA. Studies of radiation action on this DNA have been made by Fluke *et al.*,[16] by Marmur and Fluke,[17] by Ephrussi-Taylor and Latarjet,[18] and by Guild and DeFilippes.[19] A graph from the work of these last is shown as Fig. 7. The DNA was irradiated in the dry state and fast protons were used to bombard. The logarithmic relation is not simply obeyed, except at high doses. Two components are found, one with a very large inactivation volume which, when re-expressed as a molecular weight, is between 5 and 15 million, and a homogenous component of molecular weight equivalent of 300 000. The reason for the components is not clear. It may be that a fraction of the long DNA chains are so placed that they can become crosslinked by radiation action, and so cannot enter the pneumococcus to cause transformation. The smaller fraction cannot be crosslinked and is organized as polymers of an essential unit of 300 000 molecular weight, which must intercept an ionization to lose activity.

These studies again show the great sensitivity of DNA. More physical measurements have been made, notably early work by Hollaender and co-workers. More recently, Butler[20] has observed the change in viscosity of thymus DNA when irradiated by x-rays (Fig. 8). It is not plotted logarithmically, but the inactivation, as expressed by the fall in viscosity, is approximately logarithmic, and in the dry state the inactivation corresponds to an equivalent molecular weight of one million. Thus, both physical and biological measures show a strong sensitivity of DNA to ionizing radiation.

The sensitivity in aqueous solution is also high. Guild estimates that a unit of $1\frac{1}{2}$ million molecular weight can be inactivated by 100 ev of energy employed to make radicals in water.[6]

No clear proof of the precise nature of radiation action on DNA is available. Since it is so highly sensitive, and since it is a long thin molecule, the attractive hypothesis, which we adopt, is that it may be either broken, or crosslinked by ionizing radiation, whether dry or in aqueous solution. When this occurs, the result is biologically measurable. It is the biological sensitivity which is really responsible for the radiation sensitivities observed. If one chooses as the "measure" of DNA activity its ability to be digested by DNase, the radiation "molecular weight" is 4000, as was shown by Smith.[21] This is in contrast to the figure of 300 000 for transformation.

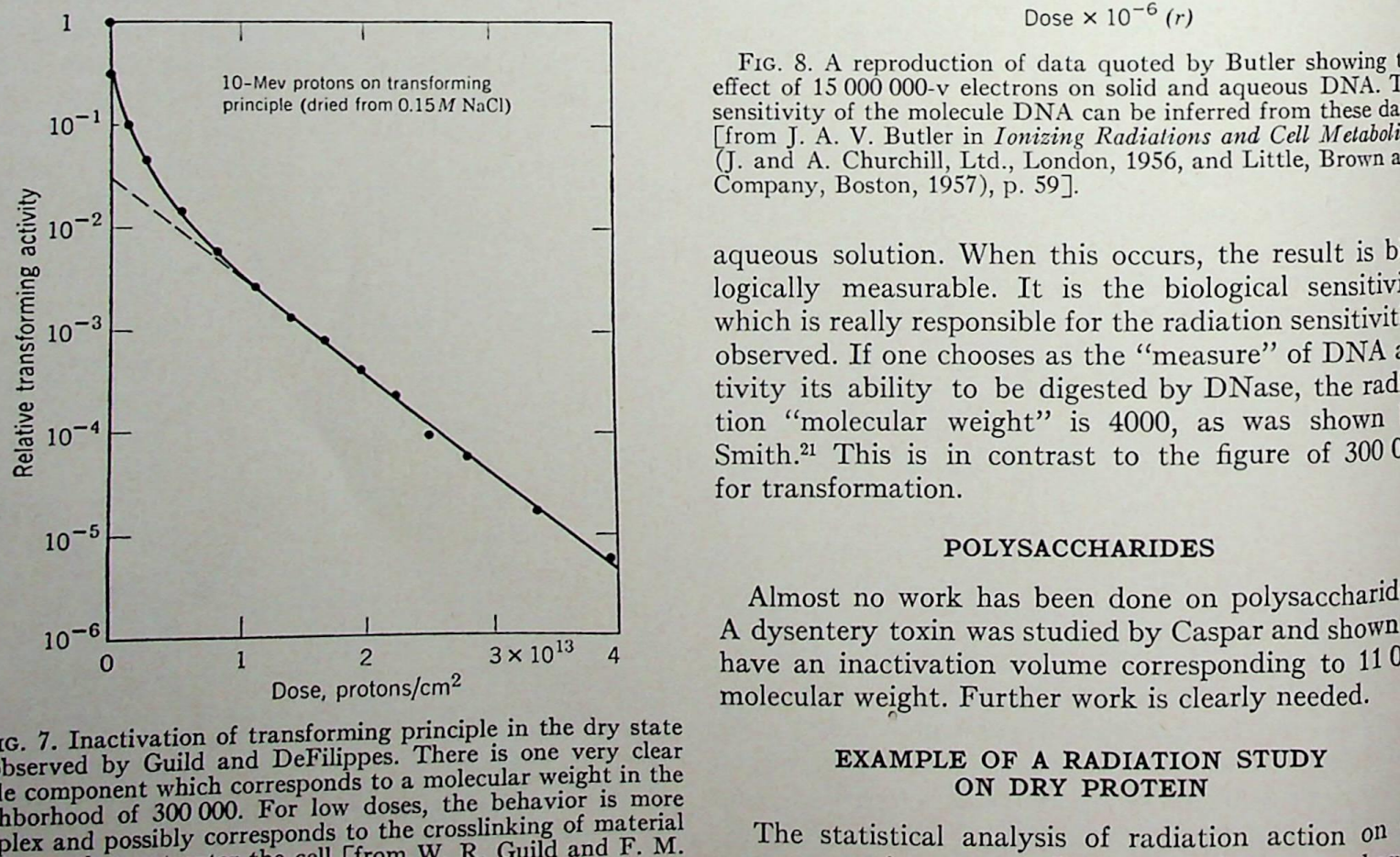


FIG. 7. Inactivation of transforming principle in the dry state as observed by Guild and DeFilippes. There is one very clear single component which corresponds to a molecular weight in the neighborhood of 300 000. For low doses, the behavior is more complex and possibly corresponds to the crosslinking of material which can then not enter the cell [from W. R. Guild and F. M. DeFilippes, Biochim. et Biophys. Acta 26, 241 (1957)].

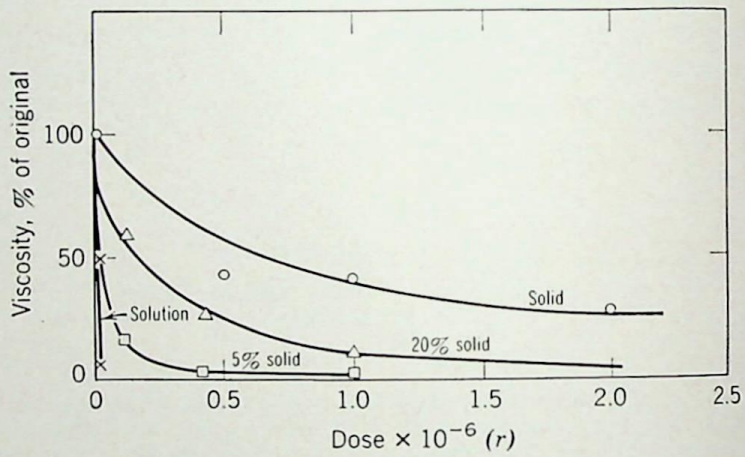


FIG. 8. A reproduction of data quoted by Butler showing the effect of 15 000 000-v electrons on solid and aqueous DNA. The sensitivity of the molecule DNA can be inferred from these data. [from J. A. V. Butler in *Ionizing Radiations and Cell Metabolism* (J. and A. Churchill, Ltd., London, 1956, and Little, Brown and Company, Boston, 1957), p. 59].

POLYSACCHARIDES

Almost no work has been done on polysaccharides. A dysentery toxin was studied by Caspar and shown to have an inactivation volume corresponding to 11 000 molecular weight. Further work is clearly needed.

EXAMPLE OF A RADIATION STUDY ON DRY PROTEIN

The statistical analysis of radiation action on an enzyme, antigen, or antibody can give more than a

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

simple estimate of a volume. As an example, it is shown in outline how an enzyme, not purified particularly, can be characterized. Since the purification and measurement by physiochemical means has to be done yet, this work offers some chance to test the predictions of such studies. The enzyme is β-galactosidase, and one aspect of the study was shown in Fig. 2 where cobalt irradiation was employed. Such radiation is random in volume, and the product IV can be found from the data, yielding V, the radiation sensitive volume. Using heavy charged particles, such as deuterons or α-particles, a fraction of their ionization, generally about 75%, is confined to narrow tracks. A fraction is spread more widely, as seen in Fig. 9 where a crude representation of the ionization produced by a 4-Mev deuteron is shown. If that fraction of the inactivation of the enzyme due to the on-track ionization can be found, then two quantities can be measured: the area S of the molecule exposed to radiation action, and the thickness E. Estimating the fractional inactivation due to the track is subject to some uncertainty, but can be done by using the Bohr

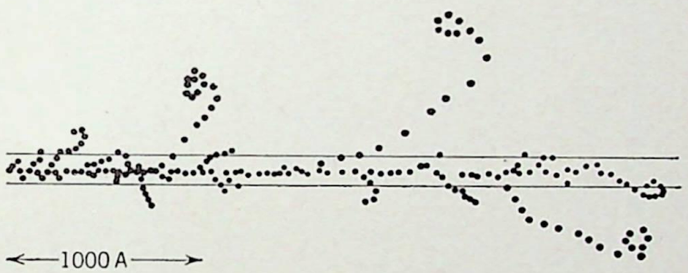


FIG. 9. Representation of the detail of a track of a deuteron. The black dots represent the ionization and the long spurs are δ-rays. In considering the part of the ionization which is along a track, allowance has to be made for the δ-rays which are off the track. The method of doing this has been worked out and permits consideration of just the part which is on the track.

theory of δ-ray production, the Bethe-Bloch energy loss expression, and some modern range measurements for slow electrons in protein.

The effect of 4-Mev deuterons on β-galactosidase is shown in Fig. 10. The data can be analyzed by a relation

$$\ln(n/n_0) = -SD, \tag{6}$$

where D is a measure of the number of deuterons per square centimeter and S is a cross section, which after correction, should bear some relation to the area of the molecule. In order to exploit the linear ionization process effectively, such cross sections can be found for different absorbers in the path of the deuteron beam. There results what is known in nuclear physics as a "Bragg curve," as seen in Fig. 11. From such a curve, the variation in cross section with the rate of energy loss of the deuteron or α-particle can be found. This is shown in Fig. 12, which also shows the result of applying the correction for the off-track ionization.

From Fig. 12, a molecular area of 8.2×10^{-13} cm^2 is deduced, and it is noted, in addition, that, for a rate of

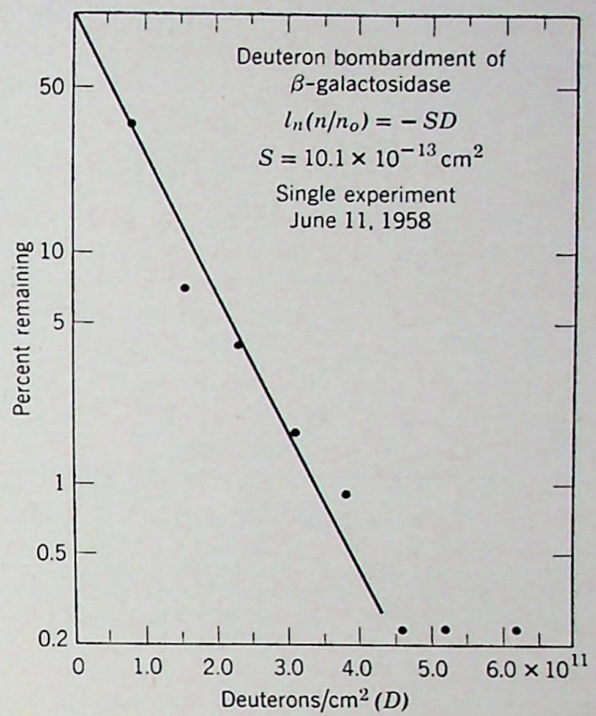


FIG. 10. The inactivation of β-galactosidase by deuteron bombardment. The relationship as indicated is obeyed and from it a target area can be inferred.

energy loss of 300 ev/100 A, the particles are 63% efficient. If this loss of efficiency is because of the fact that the ionizations are not dense enough along the track and that some "straddling" is occurring, then one should expect a relation

$$S = S_0(1-e^{-it}), \tag{7}$$

where S_0 is the maximum cross section, i is the ionization per unit track length, and t is the effective thickness. If the primary ionizations require on an average 110 ev one has, for 300 ev/100 A, $it=1$ and $i=2.7$ per 100 A, so $t=37$ A. Thus, one has the following data about

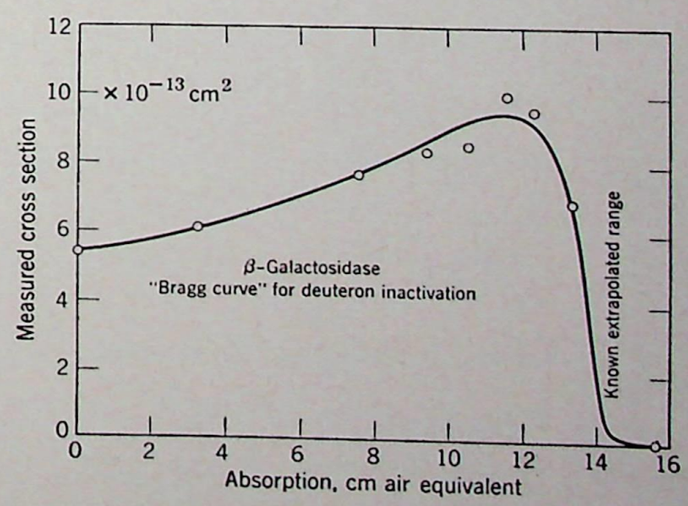


FIG. 11. A "Bragg curve" for the deuteron inactivation of β-galactosidase. The deuterons were systematically covered with foils of different thickness and their effectiveness measured. It can be seen that there is a rise as the ionization per unit path increases and then an abrupt fall as the end of the range is reached.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

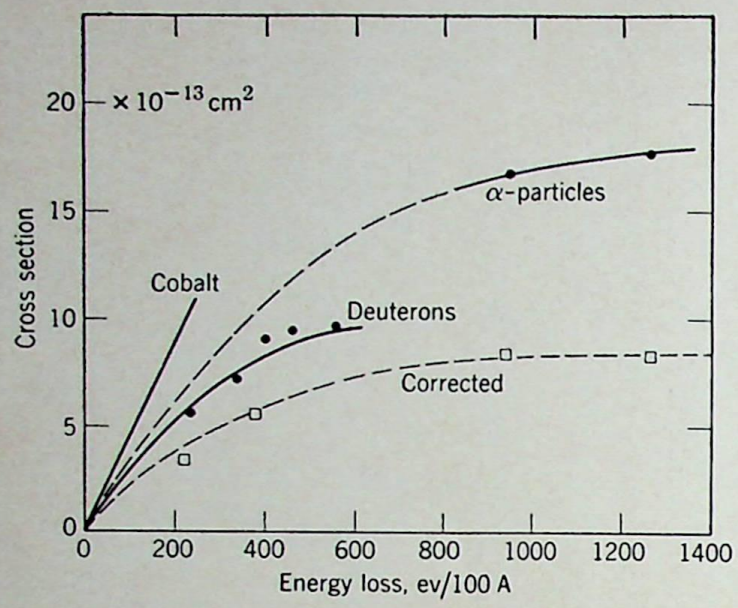


FIG. 12. The plot of the cross section *vs* rate of energy loss for deuteron and α-particle bombardment. A corrected line has been drawn in which allowance has been made for the δ-rays.

β-galactosidase:

$$\text{volume} = 4.7 \times 10^{-19} \text{ cc}$$

$$\text{area} = 8.2 \times 10^{-13} \text{ cm}^2$$

$$\text{effective thickness} = 3.7 \times 10^{-7} \text{ cm}.$$

If it is assumed that the molecule is a cylinder, radius r and length l, these can be solved for, from the volume and area, and $r = 36$ A and $l = 114$ A are obtained. The effective thickness should be $4r/\pi = 47$ A, which has to be compared with 37 A, as deduced from Fig. 12. The agreement is fairly satisfactory. Thus, the enzyme should have a molecular weight 370 000, a length of 114 A, and a diameter of 72 A, and so is roughly spherical. Such a prediction should be subject to verification.

TABLE I. Virus properties available for cross-section measurement.

Type of virus	Property	Rough results
Plant Animal Bacterial	Infectivity	Close to whole virus
Bacterial	Attachment	Very small part
Bacterial Animal	Interference	About 1/20 of virus
Bacterial	Host killing	About 1/6 of virus
Plant Animal Bacterial	Serological affinity	Very small indeed
Plant Animal Bacterial	Ability to make antibodies	Not tested
Animal Bacterial	Genetic recombination	Not tested
Some animal	Hemagglutination	Small
Bacterial Animal	Complement fixation	Not tested
Bacterial	Latent period extension	About 1/3 of virus
Some animal	Enzyme action	Small

INACTIVATION OF VIRUSES AS RELATED TO THEIR STRUCTURE

Inactivation of a whole biological system, having a structure and a number of functioning parts, presents a new problem. It is reasonably certain that radiation inactivates at least one of the molecular units of the system, but now inquiry must be made as to whether or not it has any effect on the biological functioning, or rather, any measurable effect. Viruses offer perhaps the

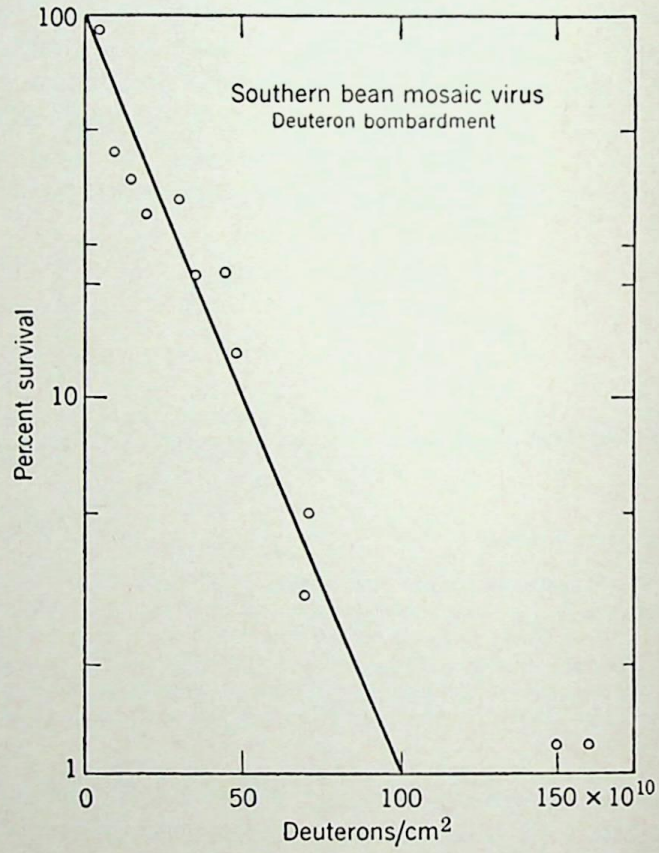


FIG. 13. Inactivation of Southern bean mosaic virus by deuterons as observed by Dimond and the author. From such data the cross section of the virus can be inferred and comes rather close to the observed cross section in electron micrographs.

the simplest system, yet a virus *is* a system and this must always be remembered. The problem can be rendered schematic as follows:

radiation → molecular inactivation → reduction of function ↓ biological consequence.

Thus, when a virus is irradiated and its ability to multiply in the host is studied, one investigates the sum total of: (1) ability to survive in outside world, (2) ability to attach to host cell, (3) ability to invade the cell, (4) ability to multiply in the cell, and (5) ability to return to outside world. Thus, bland statements

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

about "virus inactivation" should be regarded with suspicion.

Before treating separate inactivation studies, brief mention is made of two empirical correlates. The first is that the virus-inactivation cross section measured with heavy particles as bombarding agents (deuterons, α-particles) is close to the electron-micrograph cross section. This correlation fails for the large viruses but works for influenza and below. The second is that the inactivation volume measured with γ-rays is close to the volume of the nucleic-acid content. This is less well established and does not work for polio or influenza, for example, but has been rather widely adopted as a

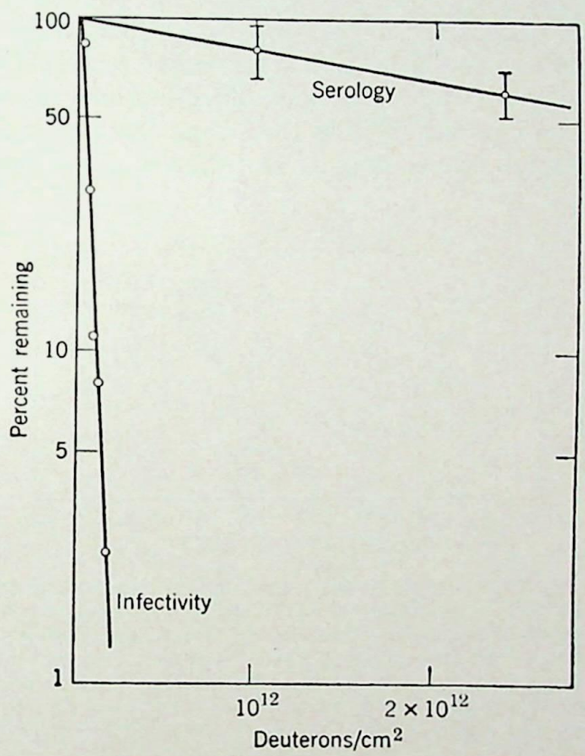


FIG. 14. Differential effect of deuterons on the infectivity and the serological combining power of tobacco mosaic virus. It can be seen that very much more radiation is needed to have any effect on the surface whereas the infectivity is very readily affected. This shows that there is a clear cut difference in the two functions of the virus and enables some estimate of the relative sizes to be made.

simple interpretation of radiation inactivation. It has some validity, but must be used with caution.

A chart is given, in Table I, of the various properties of viruses which are susceptible of study. It is clear to anyone who makes such differential radiation studies that a virus is *not* a simple molecule. In what follows, a few examples of virus studies are given, and in concluding consideration is given to what can be deduced about one virus from this kind of work. Figure 13, shows the inactivation of Southern bean mosaic virus by deuterons as measured by the ability to form local lesions on beans. The familiar relationship is found and the cross section, properly analyzed, corresponds to a

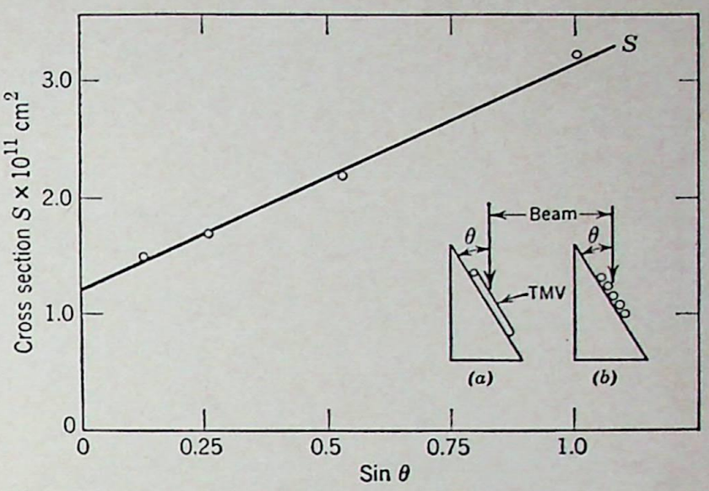


FIG. 15. Proof of the reality of irradiation target. Tobacco mosaic virus which is in long thin rods is irradiated in the manner shown in the insert of the diagram. In the first case, as the angle is varied, the target presents different aspects to the beam. This shows in the measurements as an increase in the cross section [from E. C. Pollard and G. F. Whitmore, Science **122**, 335 (1955)].

spherical particle of radius 150 A, which agrees with electron microscopy.

Figure 14 shows the effect of deuterons on dry tobacco mosaic virus (TMV) measured in terms of (a) its ability to produce local lesions, and (b) its ability to precipitate antibody to TMV. The loss of precipitating ability is

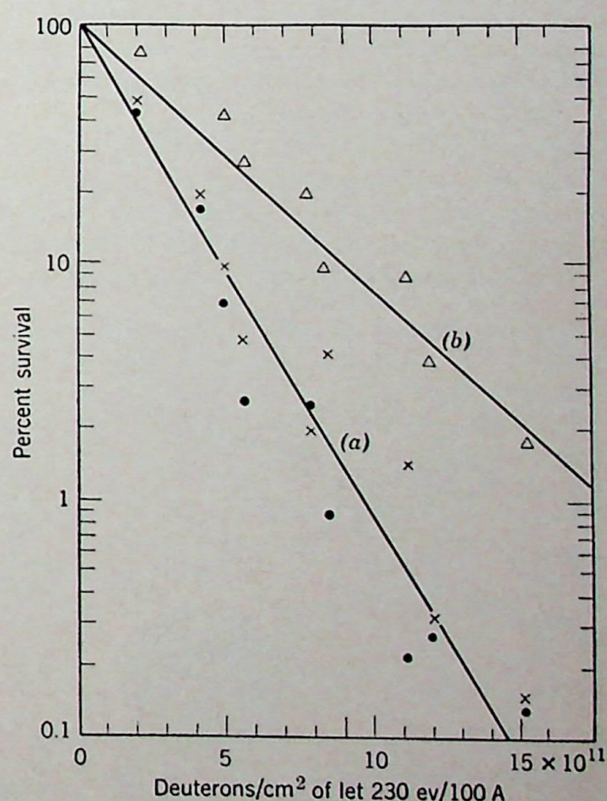


FIG. 16. Effect of deuterons on the cross reactivation of $T1$ bacterial phage as observed by Till and the author. In case (*b*), the bacterium was superinfected with undamaged virus and the fact that part of the virus material could be given from the undamaged virus to the damaged virus is shown in the lower sensitivity. The part which must be intact for cross reactivation to occur is approximately 60% of the whole virus [from J. E. Till and E. C. Pollard, Radiation Research 8, 344 (1958)].

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

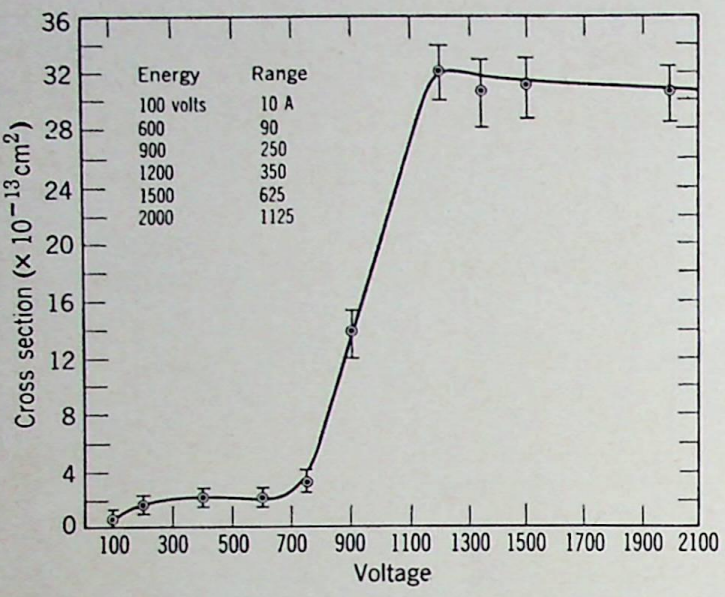


FIG. 17. The inactivation cross section of $T1$ bacteriophage with electrons of low voltage and penetration as indicated. It can be seen that, apart from a small inactivation owing presumably to some effect on the surface of the virus, the majority of the effect only occurs when the energy of the electron is sufficient to penetrate something like 250 A. This means that the sensitive part is inside the virus and also that none of it resides in the tail [from M. Davis, Arch. Biochem. Biophys. **49**, 417 (1954)].

very slight. In fact, it proved that most of the observed effect was on the virus and not on the combination with antibody. If one calculates the sensitive volume for the removal of ability to precipitate with antibody, it corresponds to a molecular weight of less than 20 000. Similar experiments on Southern bean mosaic viruses indicate that there are possibly two active sites of molecular weights 30 000 and 6000, respectively.

In view of all of this statistical reasoning, one sometimes needs more real evidence to bolster one's faith in numbers and deduction. An experiment, by Whitmore and the author,[22] on oriented samples of TMV is shown in Fig. 15. For samples of TMV held pointed at the beam, with varying angles, the cross section is least when the virus particles, which are long and thin, are held pointing at the beam and greatest when perpendicular to the beam. Controls held as at (*b*) showed no orientation effect. This experiment confirms the true physical character of the target.

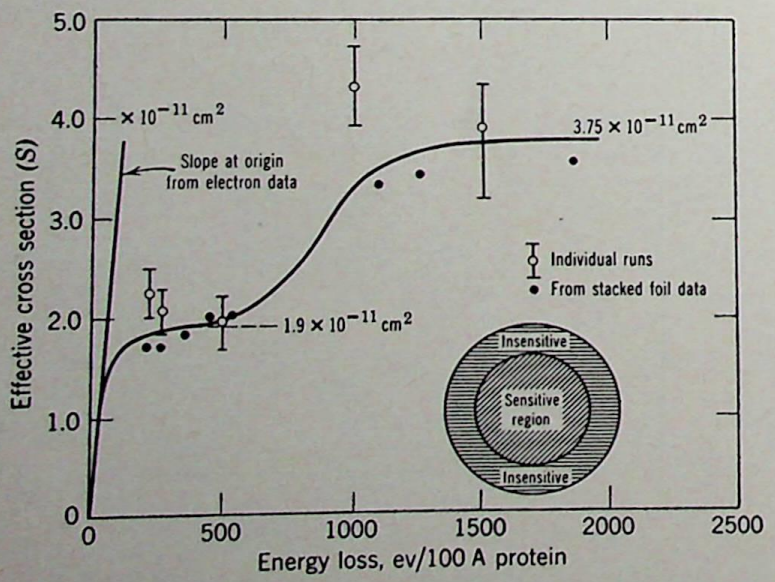


FIG. 18. Data taken by Jagger and the author on influenza virus. As the energy loss in the bombarding particle is increased, the effective cross section varies in the two-step manner. This would seem to fit with the idea that there is an inner sensitive region and an outer insensitive region and so that the virus as an internal structure [from J. Jagger and E. C. Pollard, Radiation Research 4, 1 (1956)].

BACTERIOPHAGE EXPERIMENTS

A series of phage properties has been studied. Two are selected as examples. The first is by Till and the author[23] and shows that $T1$ phage, when irradiated and allowed to enter a cell which has a second infection by a genetically different phage, can still "cross reactivate" the second phage, by which is meant that it can donate markers necessary for the progeny of both phages to function on a new host. The cross section for the loss of this ability is 60% of the single infection cross section, and it implies that 60% of the phage DNA is genetic in character while 40% is not. By quite different methods, for $T2$, Levinthal[24] has concluded somewhat similarly. The data are shown in Fig. 16.

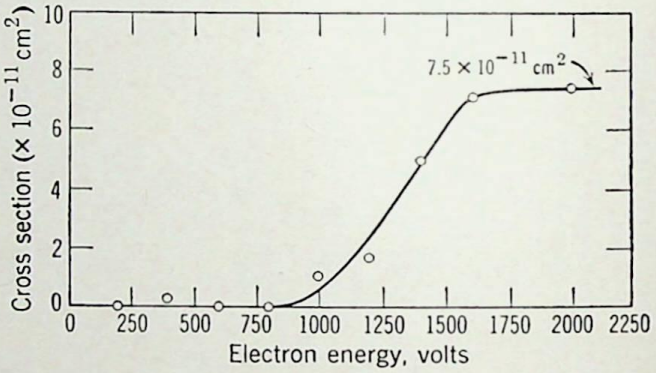


FIG. 19. Effect of low-voltage electrons on the infectivity of Newcastle disease as measured by Wilson and the author. It can be seen that there is no affect on the Newcastle disease until depth of penetration of about 200 A is observed. This means that the outer part of the virus is not concerned with the infectivity but there is a region inside which is actively concerned [from D. Wilson and E. C. Pollard, Radiation Research 8, 131 (1958)].

A technique of low-voltage irradiation, using electrons of known penetrating power, has been developed by Hutchinson *et al.*[25] Applied to $T1$ (Fig. 17), only when electrons penetrate 125 A can they produce appreciable effect on $T1$. That is, the sensitive material is imbedded in the head, and, since the tail has a thickness of only 150 A, cannot be in the tail.

NEWCASTLE DISEASE AND INFLUENZA VIRUS

Figure 18 shows an early experiment by Jagger and the author[26] on influenza virus. The infectivity in eggs was measured, and the cross section *vs* energy loss is shown. One way to analyze the data is indicated, in terms of an insensitive region surrounding one which is more sensitive. Figure 19 shows data by Wilson and the author[27] on Newcastle disease virus (NDV) using the low-voltage electron technique. Quite independent

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

evidence for the existence of a protective coat is apparent.

Figure 20 shows the assembly of conclusions about inactivation experiments on infectivity, hemagglutination, and hemolysin ability for Newcastle disease virus. Such a semispeculative figure is representative of what can be done on one virus. So far, no very contradictory evidence has come to the attention of the author which invalidates this model of the virus.

FURTHER APPLICATION AND OUTLOOK

The study of cellular structure in relation to its function by ionization inactivation can be carried beyond viruses. The effect of radiation on amino-acid uptake in *E. coli* can be analyzed to indicate that ribonucleoprotein particles, or ribosomes, are involved in the uptake. The low-voltage electron technique has been used by Preiss[28] to determine the location of the invertase in yeast cells. Currently, β-galactosidase in *E. coli* is under study and seems to be in the protoplast membrane. The technique is not harder than many others now in use and in the next few years should contribute appreciably to our knowledge of the cell.

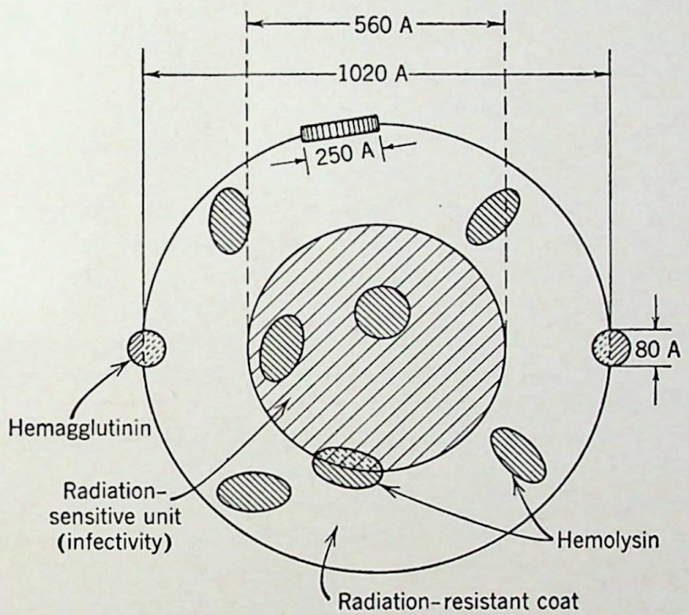


FIG. 20. Sketch of Newcastle disease virus, showing the relative size of several components. Such a drawing is schematic and serves primarily to indicate relative sizes [from D. Wilson and E. Pollard, Radiation Research 8, 131 (1958)].

BIBLIOGRAPHY

[1] J. H. Northrop, J. Gen. Physiol. **17**, 359 (1934).
[2] R. G. Hussey and W. R. Thompson, J. Gen. Physiol. **5**, 647 (1923).
[3] F. Dessauer, Z. Physik **12**, 38 (1923).
[4] J. A. Crowther, Proc. Roy. Soc. (London) **B96**, 207 (1924).
[5] E. U. Condon and H. M. Terrill, Cancer Research **11**, 324 (1927).
[6] F. M. DeFilippes and W. R. Guild, Radiation Research (to be published).
[7] F. Hutchinson and D. A. Ross, Radiation Research (to be published).
[8] C. L. Smith, Arch. Biochem. Biophys. **50**, 322 (1954).
[9] D. Lea, K. M. Smith, B. Holmes, and R. Markham, Parasitology **36**, 110 (1944).
[10] D. J. Fluke and E. C. Pollard, Ann. N. Y. Acad. Sci. **59**, 484 (1955).
[11] E. C. Pollard, National Research Council Nuclear Science Series, Report No. 17, 1 (1954).
[12] W. Gordy, W. B. Ard, and H. Shields, Proc. Natl. Acad. Sci. U. S. **41**, 983 (1955).
[13] R. Braams, F. Hutchinson, and D. Ray, Nature (to be published).
[14] F. Hutchinson, Radiation Research **7**, 473 (1957).
[15] P. Alexander, Radiation Research **6**, 653 (1957).
[16] D. Fluke, R. Drew, and E. Pollard, Proc. Natl. Acad. Sci. U. S. **38**, 180 (1952).
[17] J. Marmur and D. J. Fluke, Arch. Biochem. Biophys. **57**, 506 (1955).
[18] H. Ephrussi-Taylor and R. Latarjet, Biochim. et Biophys. Acta **16**, 183 (1955).
[19] W. R. Guild and F. M. DeFilippes, Biochim. et Biophys. Acta **26**, 241 (1957).
[20] J. A. V. Butler in *Ionizing Radiations and Cell Metabolism* (J. and A. Churchill, Ltd., London, 1956, and Little, Brown and Company, Boston, 1957), p. 59.
[21] C. L. Smith, Arch. Biochem. Biophys. **45**, 83 (1953).
[22] E. C. Pollard and G. F. Whitmore, Science **122**, 335 (1955).
[23] J. E. Till and E. C. Pollard, Radiation Research **8**, 344 (1958).
[24] C. Levinthal, Proc. Natl. Acad. Sci. U. S. **42**, 394 (1956).
[25] F. Hutchinson, Ann. N. Y. Acad. Sci. **59**, 494 (1955).
[26] J. Jagger and E. C. Pollard, Radiation Research **4**, 1 (1956).
[27] D. Wilson and E. C. Pollard, Radiation Research **8**, 131 (1958).
[28] J. W. Preiss, Arch. Biochem. Biophys. (to be published).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

34
Some Aspects of Cellular Radiobiology*

T. H. Wood

Department of Physics, The University of Pennsylvania, Philadelphia, Pennsylvania

In a preceding paper, Bloom (p. 21) describes experiments in which a stream of protons was collimated into a microbeam a few microns in diameter which could be used selectively to irradiate and inactivate chosen regions of living cells. Such studies should become increasingly important in modern biology as they allow correlations to be made between structure and function. However, before radiation probes can be used to give meaningful answers to such questions, it is necessary that one learn more about those events intervening between the original radiation insult and the observable end effect being assayed. In the complicated cellular systems discussed in the following, lack of information about these intervening processes is almost complete and few unambiguous answers are available.

Considered are two of the primary problems of cellular radiobiology. The first of these—With over-all cellular irradiation, what are the sites of damage as scored by various end-effects?—is a question which, in many cases at least, is answered. The second problem depends upon the answer to the first and is: Must the energy initiating the processes leading to the observable effect be laid down by the ionizing radiation particles directly within the target sites or can a portion of this activation energy migrate from the environment to the sensitive sites?

With the task at hand defined in rather broad terms, I should like to limit it by considering almost exclusively a particular microorganism (the yeast *Saccharomyces cerevisiae*) and by dealing with only one end point of radiation damage, namely, the loss of ability of an irradiated cell to produce a visible colony. Although one could use other end points such as particular genetic mutations, division delay, biochemical abnormalities, changes in cellular permeability, etc., the loss of ability to reproduce is, from a teleological point of view, of greatest importance.

Bloom, in his description of the microbeam experiments, has already given the clue that, in his test system, the nucleus is much more sensitive to irradiation than is the cytoplasm. So, the first question can be rephrased: What is the relative radiosensitivity of the nucleus and the cytoplasm in yeast as scored by colony formation?

It is possible to derive from a single strain of yeast related strains in which the chromosomal material is present in quantities which are multiples of a basic unit; such strains are designated as haploid, diploid, triploid, etc., up to hexaploid.[1] It has been most profitable to compare the radiosensitivities of related yeasts of various ploidies. In a typical experiment, yeast cells are suspended in an aqueous environment under various conditions and irradiated. Aliquots are then plated on nutrient agar on which a surviving cell, by our criterion, will produce a visible colony (about a million cells). The survival of such an irradiated population can then be compared with that of an unirradiated one. A semilogarithmic plot of the surviving fraction of haploid yeast as a function of dose is shown in Fig. 1. Exponential survival is obtained, the slope of the survival curve being a useful index of the radiosensitivity. The inactivation kinetics are first order, indicating that the inactivation process is caused by a single event, presumably an energy transfer to a single molecule. One might now ask the question: What single molecule is necessary for the continued functioning of the cell? One is led almost inevitably to the conclusion that the pertinent molecules must be molecules of the genetic apparatus of the cell.

The x-ray survival curve of diploid yeast is shown in Fig. 2.[2] Two points are noteworthy: (1) these cells are much more radioresistant than are haploid cells, and (2) the inactivation kinetics are of higher order than first. The first attempt to explain mechanistically the comparative response of haploid and diploid strains was that of Zirkle and Tobias.[3] They assumed that radiation damage in this system is owing to the induction of recessive lethal mutations. This is diagram-

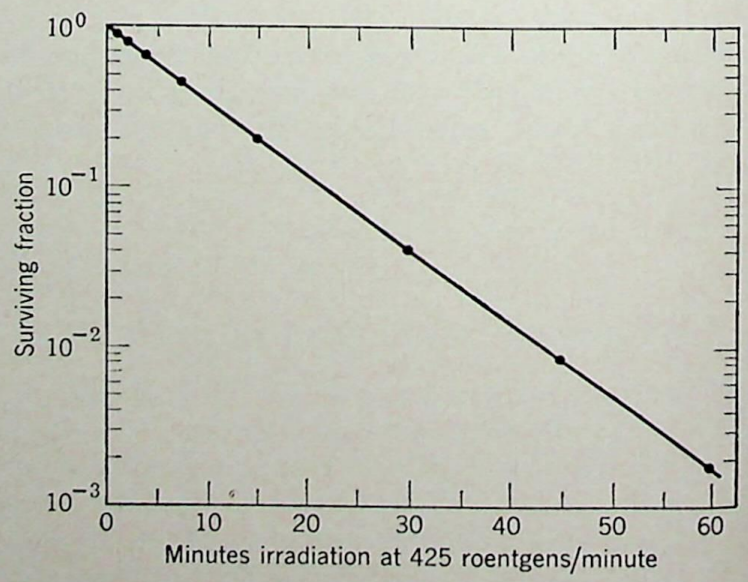


FIG. 1. X-ray survival curve of haploid yeast. Experimental points have errors of 2% or less (approximately the size of the points).

* A portion of the work reported herein was supported by a research grant (C–2390) from the National Cancer Institute, National Institutes of Health, Public Health Service, U. S. Department of Health, Education, and Welfare.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

matically represented in Fig. 3 in which all of the chromosomal material of the haploid is shown in a continuous array; this will be duplicated in the diploid and will be present m-fold in yeast of ploidy m. To successfully inactivate a multiploid cell, corresponding sites on homologous chromosomes such as c, c', c'', etc., must be simultaneously inactivated. Such a model can successfully correlate haploid and diploid survival and predicts increasing radioresistance for yeast of higher ploidy.

To test this hypothesis, Mortimer has extended these studies to yeast of ploidies up to six and finds that the triploid, tetraploid, pentaploid, and hexaploid are progressively *less* radioresistant, in direct contradiction to the predictions of the recessive lethal model.[1] Figure 4 shows that the doses required to produce the same

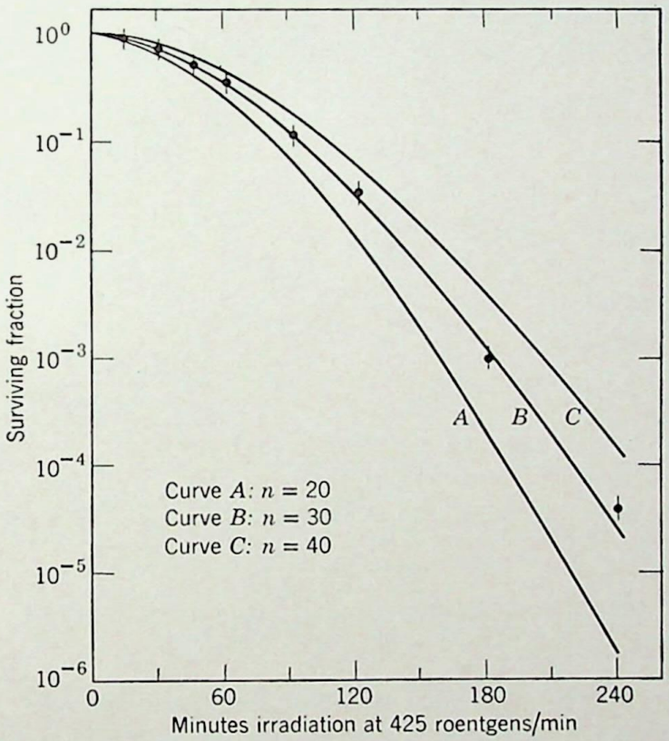


FIG. 2. X-ray survival curve of diploid yeast.

percentage inactivation decrease inversely with the ploidy for ploidies above haploid.[4]

In another series of elegant experiments, Mortimer has shown that x-ray inactivation of haploid yeast may be owing to the induction of either recessive or of dominant lethal mutations.[5] By micromanipulation, individual irradiated cells of various ploidies are mated with unirradiated cells. A diagram of such an experiment with haploid cells is shown in Fig. 5. Large numbers of cells have to be mated as such interpretations must be based on statistical evidence. Operationally, recessive lethal mutations in haploid cells are defined as those not causing loss of ability to divide when mating with an undamaged haploid cell; dominant lethals are defined as those causing loss of ability to divide.

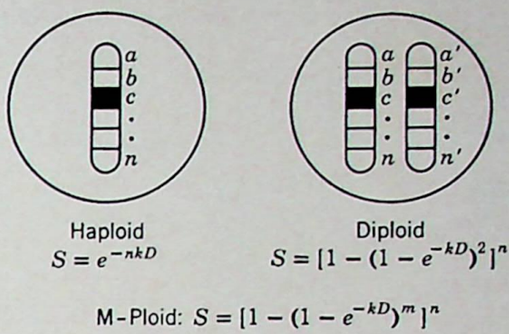


FIG. 3. Diagram for recessive lethal model.

Dominant lethals would be expected to be proportional to the total chromosome length, that is, to the ploidy. The data of Fig. 4 indicate that with yeast cells of higher ploidy the radioresistance is inversely related to the ploidy, or, the radiosensitivity is directly proportional to the ploidy. Thus, with yeasts of ploidy higher than diploid, the predominant path of inactivation as scored by colony formation is interpreted by Mortimer as owing to the induction of dominant lethal mutations.

It is very difficult to relate this dependence of radiosensitivity on ploidy to cytoplasmic damage resulting from irradiation. Cell volume is proportional to ploidy[1] and the diploid is more radioresistant than the haploid. Thus, if radiation damage were the result of cytoplasmic effects, the radioresistance should continue to increase with ploidies above two in contradiction to the observed results.

Additional evidence that the predominant path of radiation damage is by means of genetic damage is afforded by Tobias' experiments with defective diploid cells in which recessive radiation damage is propagated to the daughters of irradiated cells (p. 289).

In general, then, with yeast loss of ability to form

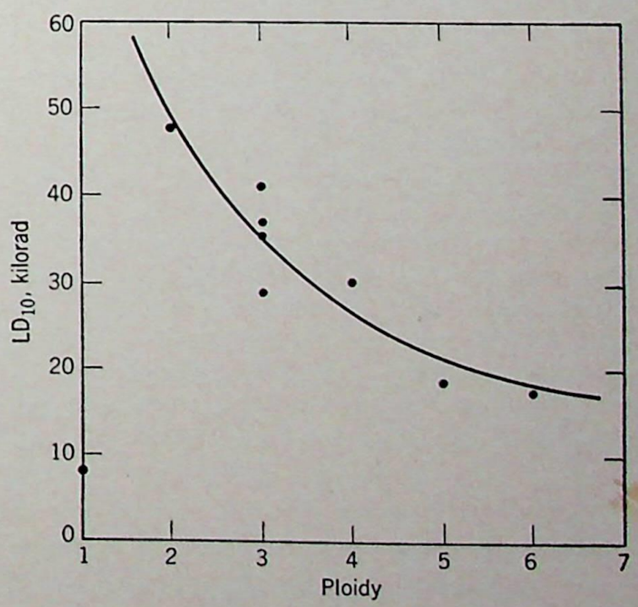


FIG. 4. Radioresistance of yeast as a function of ploidy [from R. K. Mortimer, University of California Radiation Laboratory Rept. 3902 (August, 1957)].

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

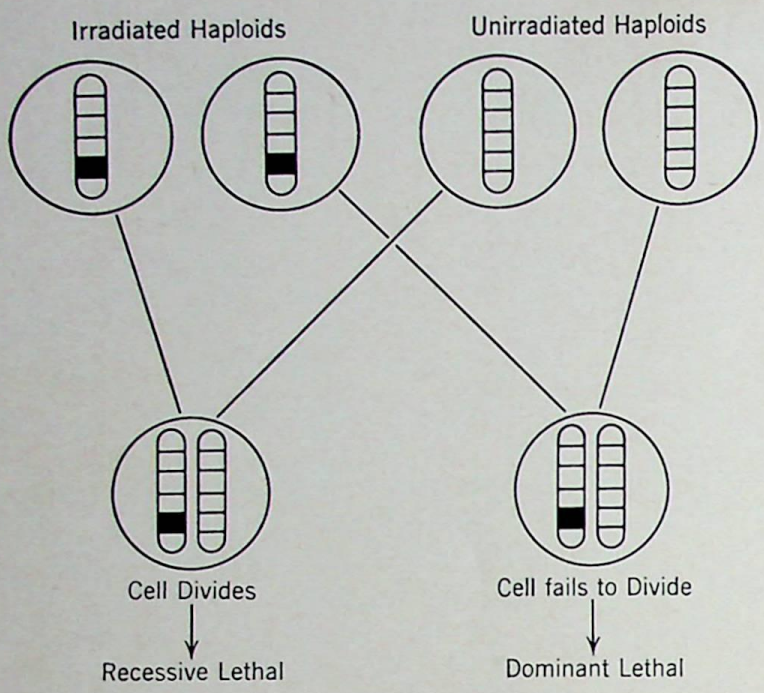


FIG. 5. Operational scheme for differentiating between recessive and dominant lethal mutations induced by x-rays in yeast.

colonies following x-irradiation can be explained in terms of genetic effects. The microbeam experiments of Zirkle and Bloom,[6] experiments with eccentrically located nuclei,[7,8] and many other experiments all indicate that nuclear inactivation is of primary importance in many types of cells. The relatively small number of cases to the contrary that have been reported (e.g., Duryee with frog eggs,[9] Bacq *et al.* with alga[10]) are interesting exceptions to this rule.

Turn now to the question of whether the radiation energy causing cellular inactivation is laid down directly in the molecules whose modification causes cellular inactivation or whether a portion of this energy can migrate from the molecular environment to the affected sites. Pollard (p. 273) has outlined in some detail the processes of direct-action inactivation in dried materials. There can be no doubt that such inactivation plays a role in cellular radiobiology. However, classical direct-action theory in its simplest form is unable to account for various experimental results, in particular those concerned with modification of radiation effects by various environmental factors such as oxygen concentration and phase state.[11] Furthermore, with the development of radiation chemistry and the demonstration of *in vitro* inactivation of various cellular constituents by free radicals produced by the irradiation of water, it became necessary to consider the importance of such processes in cellular systems. There thus evolved the indirect action hypothesis in which free radicals produced by the irradiation of water were postulated to bring about at least some of the initial radiation damage.

The most successful formulation of these concepts

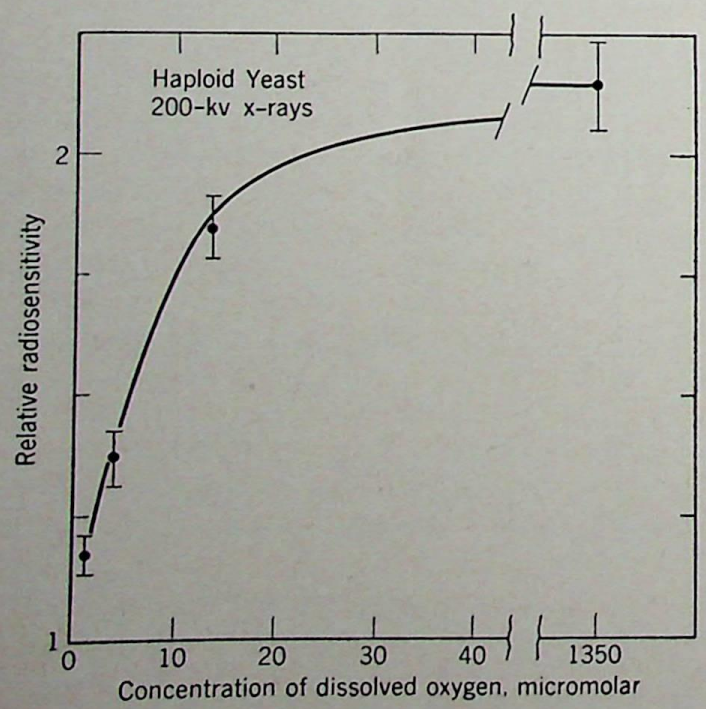


FIG. 6. Variation with oxygen concentration of x-ray sensitivity of haploid yeast [from P. Howard-Flanders and T. Alper, Radiation Research 7, 518 (1957)].

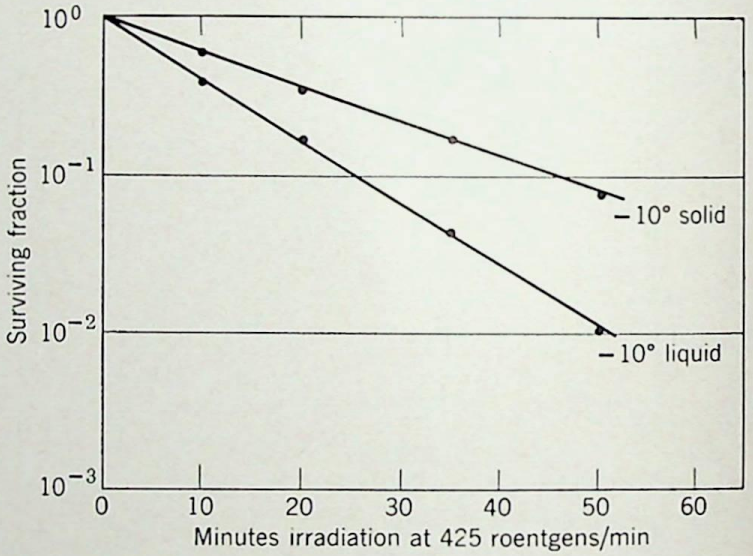


FIG. 7. X-ray survival curves of haploid yeast irradiated at −10°C, liquid and solid phases.

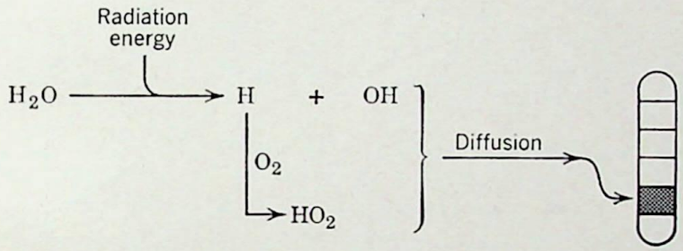
was in the migration model of Zirkle and Tobias.[3] In this model, a proper averaging was made of the contributions to inactivation by various species of free radicals, produced at various distances from the target molecules, under various experimental conditions. A single effective collision between an energy carrier and a critical molecule was postulated to be effective in bringing about molecular change in conformity with the first-order inactivation kinetics experimentally observed in many of the simple unicellular systems. This model also includes as a part of itself the direct-action type of inactivation.

One of the most studied of the modifying parameters is the oxygen concentration in the cellular suspension during irradiation. Haploid yeast cells irradiated in the absence of oxygen are approximately twice as resistant as those irradiated under aerobic conditions (Fig. 6[12]). The oxygen effect is operative at oxygen tensions as low as 10 micromolar. A comparable oxygen effect has been found for a tremendous variety of materials using

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

a considerable number of end results. Many chemicals have been found to decrease the radiosensitivity, but it is not yet certain that such protection acts other than by modifying the existing oxygen concentration. The phase state of the test system, whether frozen or liquid, greatly modifies radiosensitivity.[13] In Fig. 7 are shown survival curves for haploid yeast irradiated at −10°C in the liquid phase (supercooled) and in the frozen state.

Until quite recently, the modification of radiation response by such environmental factors as oxygen concentration, various chemicals, or phase-state change was considered as evidence that at least the modifiable portion of the radiation effect was mediated through indirect-action mechanisms. A simple formulation of indirect action is diagrammed in Fig. 8. Oxygen modification can be explained easily in this model by the elimination of those free radicals depending on the presence of oxygen for their formation. Modification by chemicals could be owing to competition for the free



Modification by:
Oxygen: Elimination of HO_2
Chemicals: Competition for radicals
Shielding of targets
Competition for oxygen
Phase State: Decreased diffusion in solid state
or
Dehydration Decreased number of free radicals produced near targets

FIG. 8. Diagram of simplified indirect-action mechanisms.

radicals between the added chemicals and the target sites, to diminution of the oxygen effect, or to masking of sensitive molecular side chains (e.g., sulfhydryl groups). The phase-state change would operate by either removing water from availability for these reactions or by greatly decreasing the diffusion of free radicals. On the other hand, such modifying agents would not be expected to greatly modify the ionization levels of the target molecules and modification is difficult to explain using the concepts of classical direct action.

Alper, Howard-Flanders, and Alexander have questioned the assumption that modification is *a priori* evidence that an indirect action mechanism is operative.[12,14,15] For example, Alexander has shown that crystalline trypsin also shows an oxygen effect.[14] However, this evidence is not conclusive as this system still contains about 5% water which might play an important role in the oxygen effect. They suggest that the

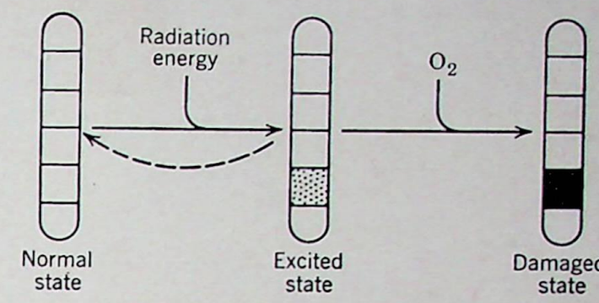


FIG. 9. Diagram of oxygen modification of radiosensitivity

oxygen effect is due to reaction of an excited state with oxygen to produce an irreversibly damaged site (Fig. 9). Recently, Hutchinson has interpreted his results on the *in vivo* inactivation of yeast enzymes by x-rays in terms of the migration model and has found that a diffusion distance for the pertinent inactivating free radicals of about 30 A is involved, a distance an order of magnitude less than had previously been estimated.[16] At the lowest oxygen concentrations at which the oxygen effect is important, the distance between oxygen molecules is about 700A. It is, therefore, unlikely in Hutchinson's system that there would be opportunity of reaction between oxygen and hydrogen free radicals produced by irradiation to give rise to the hydroperoxyl free radical ($H+O_2=HO_2$), the generally assumed mechanism for the oxygen effect. Other objections to the hydroperoxyl free radical as an important factor in cellular radiobiology have been discussed elsewhere in detail.[12] Thus, while modification of radiation damage in the truly dried state has not been demonstrated, the generally accepted mechanism for the oxygen effect via the hydroperoxyl free radical is subject to question. However, the mechanism diagramed in Fig. 9 could also be operative for indirect action events, and at the present time this seems to be the best operational hypothesis for the oxygen effect. Recently, estimates have been made on the lifetime of the excited states that may be involved in the oxygen effect. At the low oxygen concentrations necessary for the oxygen effect, there are approximately 10^6 collisions/sec between oxygen molecules (assuming that they are free to diffuse within the cell) and a hypothetical target having a molecular weight of 100. Thus, it is necessary to assume that the excited state produced by the radiation energy must have a lifetime of at least a microsecond in length. Work by Howard-Flanders and Moore with bacteria in which cells irradiated under either aerobic or anaerobic conditions were quickly switched to the opposite condition indicates that the lifetime of the oxygen-sensitive excited state is less than 20 msec.[17]

Howard-Flanders has also found that nitric oxide can be used to mimic oxygen.[18] He suggests that the action of oxygen and nitric oxide in increasing radiosensitivity may be associated with the fact that both have unpaired electrons in π orbitals.

The change in radiosensitivity associated with the

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

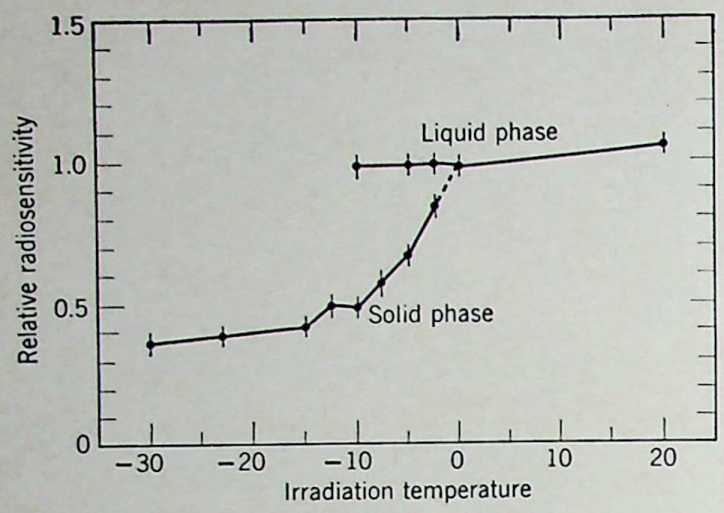


FIG. 10. Dependence of radiosensitivity on irradiation temperature: −30°C to 0°C, liquid and solid phases.

phase-state change from liquid to frozen cells has been mentioned already. Figure 10 shows the radiosensitivity over a temperature region down to −35°C in both the liquid and solid phase.[13] The progressive decrease in radiosensitivity with irradiation temperatures below 0°C suggested to us that the decrease might be associated with a progressive freezing of cellular water. To test this hypothesis, microcalorimetric studies were performed on frozen yeast suspensions. Figure 11 shows that cellular water is progressively frozen down to about −20°C, this region of constant freezing correlating nicely with the region of constant radiosensitivity. A certain fraction of the cellular water, 9%, is nonfreezable; this is one definition of bound water.[19]

A. M. Rosenberg, in our laboratory, has developed an interesting way to control effectively the amount of free water within yeast.[20] Very high concentrations of various chemicals (ethanol, glycerol, glucose, etc.) can be used to either dehydrate the cell or to tie up cellular water which may play a role in radiobiological processes. Figure 12 shows the progressive decrease in radiosensitivity with increase in glycerol concentration. The radioresistance increases by about a factor of four with 6.9 M glycerol. Additional protection is afforded by anoxic conditions, the maximum radioresistance being about 5 times that typical of aerobic cells in standard suspensions. We have been able to verify by several methods that glucose, for example, dehydrates the cell approximately to the degree expected from simple theory. Figure 13 summarizes the points mentioned in the foregoing—namely, that there is a correlation between the residual fractions of radiosensitivity and of unfrozen cellular water with respect to freezing temperature, and between the residual fractions of radiosensitivity and of cellular volume with respect to osmotic concentration. Furthermore, these two types of effective water removal are apparently closely related.

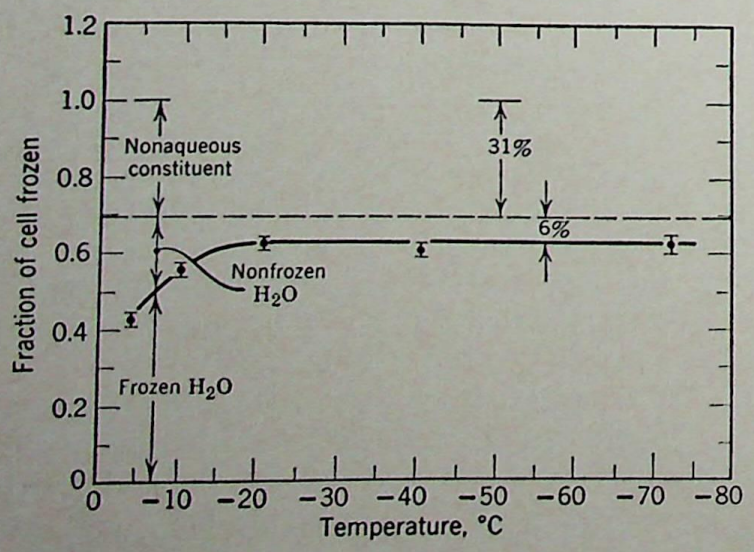


FIG. 11. Fraction of the normal yeast cell frozen as a function of temperature.

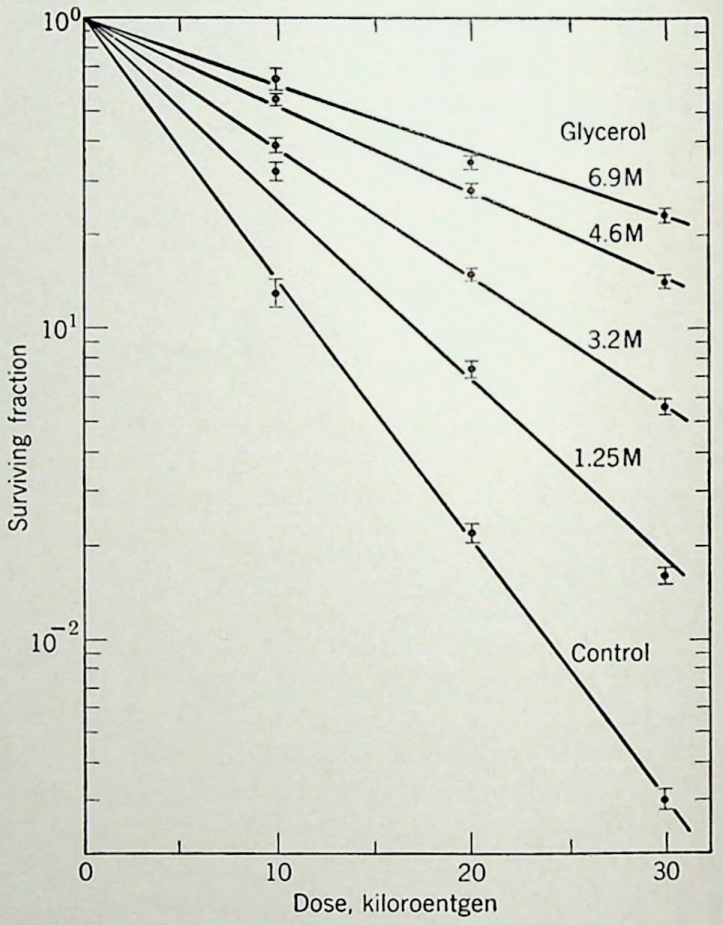


FIG. 12. X-ray survival curves for haploid yeast irradiated in different concentrations of glycerol.

From the results in Figs. 10 and 11, one might expect that the temperature dependence of yeast would not greatly change for irradiation temperatures below −35°C. However, when yeast suspensions are frozen at a rate characteristic of the irradiation temperature, an apparently anomalous behavior is observed below −35°C, the radiosensitivity progressively increasing in the region from −35°C to −72°C where a sensitivity is observed approximately equal to that typical of the liquid phase (Fig. 14[21]). This behavior can be explained by an examination of the physical processes that occur within the cell during freezing. Meryman and Platt[22] have shown that, in mammalian liver cells, the rate of freezing (the speed at which the ice-water boundary advances) is of great importance in determining whether

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

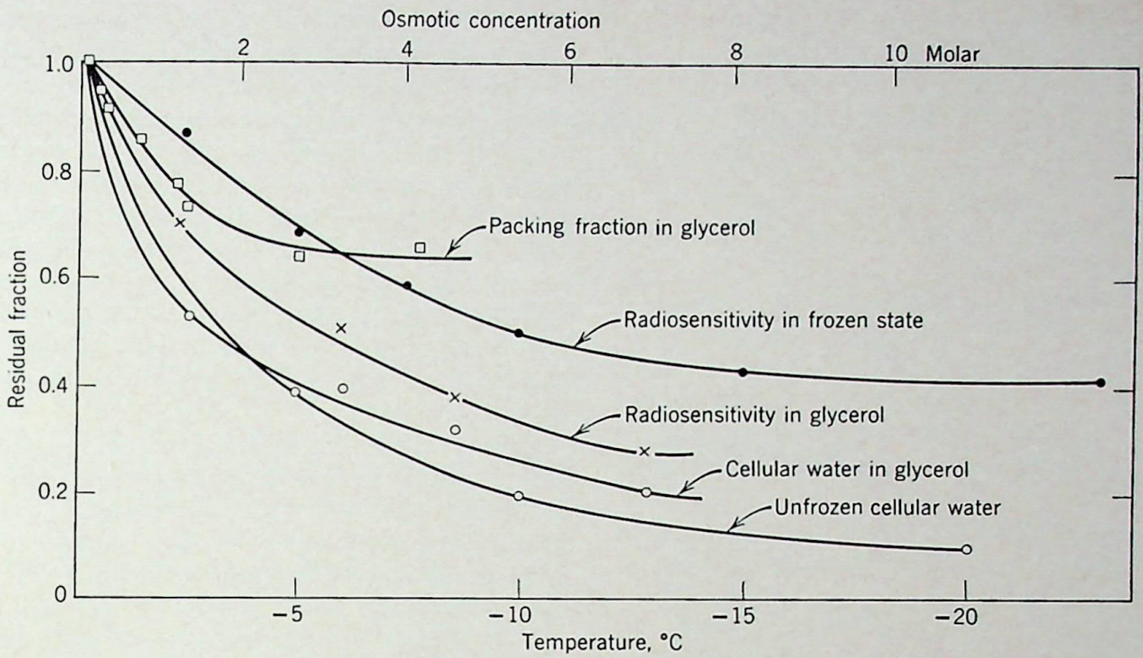


FIG. 13. Relationship of the radiosensitivity of yeast to the loss of cellular water. Glycerol concentrations are to be read at the uppermost abscissa, freezing temperatures at the bottom. The curve labeled "packing fraction in glycerol" shows the relative volume occupied by a standard amount of yeast after centrifugation in various concentrations of glycerol.

the cellular water is frozen within or without the cell. At low rates of freezing (less than 1.5 mm/min), the increase in external osmotic pressure owing to the freezing of the external medium causes water to move outward across the cell membrane where it is then frozen. Thus, the cell is effectively dehydrated. At high rates of freezing (greater than 1.5 mm/min), however, the cellular water freezes within the cell before the increase in external osmotic pressure is effective in dehydrating the cell. In the yeast system used here, this critical velocity of freezing (1.5 mm/min) occurs at −35°C; thus, for cells frozen at temperatures above −35°C, dehydration would occur. For freezing at temperatures below −35°C, some internal freezing of water would occur before dehydration would be complete. These mechanisms are diagramed in Fig. 15.

The results in Fig. 14 are interpretable in light of the migration model using the foregoing concepts of cellular freezing. On irradiation, free radicals are formed in all of the cellular water, whether liquid or frozen, whether within or without the cell, approximately in equal number per unit volume of water. We further assume that free radicals (or their products) produced by the irradiation of the ice are largely trapped within the ice crystal. The investigations of Stewart and Ghormley[23] give credence to this assumption. Now, if the frozen cellular water is largely exterior to the cell, the free radicals liberated on thawing will be effectively diluted by the external medium and the probability of an effective interaction between them and the sensitive regions of the cell will materially decrease. On the other hand, if a significant amount of the cellular water is frozen within the cell, the free radicals released on thawing will be in much the same positions they would have been in on irradiation of the unfrozen cell. Thus, if the radiobiologically significant cellular water is frozen within the cell, the radiosensitivity under these conditions would be not dissimilar to that typical of the liquid phase.

There are three critical tests for the foregoing hypothesis. If radiation sensitivity is primarily dependent on the degree of external dehydration rather than on the absolute amount of cellular water frozen (i.e., the temperature), then slow freezing to temperatures lower than −35°C by stepwise freezing at −5°C, −10°C, −20°C, etc. should dehydrate the cells to the same degree as freezing at −35°C. The resulting radiosensitivity should be the same as that obtained with cells frozen at −35°C; from Fig. 14 (slow O_2 curve) this can be seen to be the case. Secondly, if fast freezing at

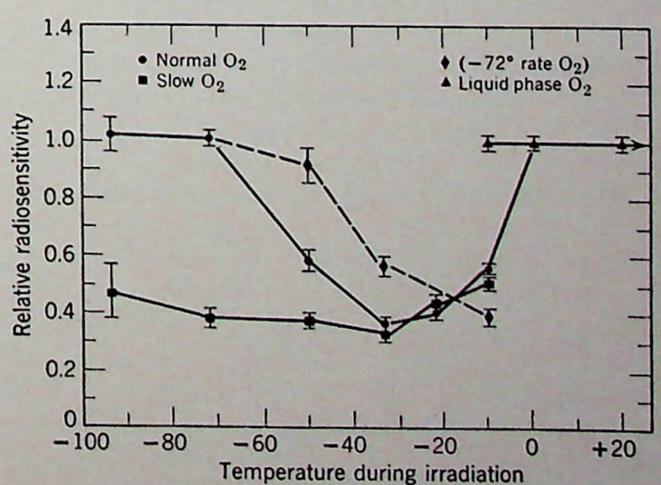


FIG. 14. Relative radiosensitivity of haploid yeast as a function of temperature, phase state, and rate of freezing.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

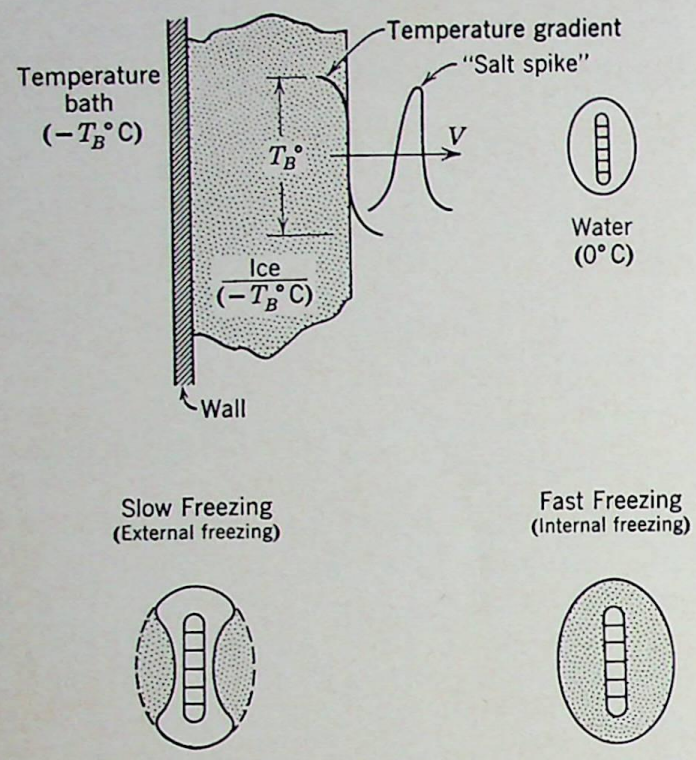


FIG. 15. Mechanisms for fast and slow freezing in yeast.

−72°C internally freezes the radiobiologically significant water, then freezing at even lower temperatures, say, −95°C, should produce the same radiosensitivity as that obtained at −72°C and in the liquid phase; this was also found to be the case. Thirdly, if fast freezing at −72°C freezes the radiobiologically significant water within the cell, then subsequent x-irradiation of such cells at temperatures higher than −72°C should produce the same effect as that obtained at −72°C. This is true at −50°C, but at higher temperatures there is a progressive loss of radiosensitivity. This decrease in sensitivity with increasing temperature for these prefrozen cells is qualitatively similar to the rate of disappearance of hydrogen peroxide produced by irradiation of ice that has been observed by Stewart and Ghormley and is probably the result of back reactions among free radicals not firmly trapped in the ice.[23]

No successful explanation of the phase effect in radiobiology based on the direct action model has been advanced (see Wood[24] for review). Probably the best way to differentiate between direct-action mechanisms and indirect-action mechanisms is by use of the phase effect. In the haploid-yeast system, 60% of the radiation damage under either aerobic or anaerobic conditions is capable of modification by the change in phase from liquid to solid state. Thus, if the above interpretations are correct, no more than 40% of the over-all radiation damage is owing to direct action. If oxygen modification operates through an indirect-action mechanism, then our results both with frozen, anoxic cells and with glycerol-dehydrated, anoxic cells indicate that at least 80% of the x-ray inactivation of this system is owing to indirect action.

In summarizing, it is difficult to make any broad statements that have general applicability for cellular radiobiology. The modification by oxygen of almost all radiobiological responses studied indicates the necessity of understanding the associated mechanisms. The generally accepted hydroperoxyl-radical mechanism is certainly questionable. At the present, the best operational tool to differentiate between direct and indirect types of inactivation is the change in radiosensitivity with phase state. The success of the migration model in dealing with modification of radiosensitivity by various physical and biological parameters is most encouraging and this model, even though its details may need revision, has been most valuable in suggesting new experiments.

In studying a biological unit as complicated as a cell, it is desirable to approach the problem from both the molecular level and from the gross phenomenological level. Phenomenological studies similar to those described here have greatly profited from work done with simpler systems such as those described by Pollard (p. 273). On the other hand, work with more complicated cellular systems has given a direction to molecular studies and has focused attention on the necessity for working with materials in an aqueous environment It is to be hoped that it will not be too long before these two types of studies have overlapped.

BIBLIOGRAPHY

[1] R. K. Mortimer, Radiation Research **9**, 312 (1958).
[2] T. H. Wood, Proc. Soc. Exptl. Biol. Med. **84**, 446 (1953).
[3] R. E. Zirkle and C. A. Tobias, Arch. Biochem. Biophys. **47**, 282 (1953).
[4] R. K. Mortimer, University of California Radiation Laboratory Rept. 3902 (August, 1957).
[5] R. K. Mortimer, Radiation Research **2**, 361 (1955).
[6] R. E. Zirkle and W. Bloom, Science **117**, 487 (1953).
[7] R. C. von Borstel and H. Moser in *Progress in Radiobiology*, J. S. Mitchell, B. E. Holmes, and C. L. Smith, editors (Oliver and Boyd, London, 1956), p. 211.
[8] R. E. Zirkle, J. Cellular Comp. Physiol. **2**, 251 (1932).
[9] W. R. Duryee, J. Natl. Cancer Inst. **10**, 735 (1949).
[10] Z. M. Bacq, F. Vanderhaeghe, J. Damblon, M. Errera, and A. Herve, Exptl. Cell Research **12**, 639 (1957).
[11] D. E. Lea, *Actions on Radiations on Living Cells* (Cambridge University Press, London, 1955), revised edition.
[12] P. Howard-Flanders and T. Alper, Radiation Research **7**, 518 (1957).
[13] T. H. Wood, Arch. Biochem. Biophys. **52**, 157 (1954).
[14] P. Alexander, Radiation Research **6**, 653 (1957).
[15] T. Alper, Radiation Research **5**, 573 (1956).
[16] F. Hutchinson, Radiation Research **7**, 473 (1957).
[17] P. Howard-Flanders and D. Moore, Radiation Research **9**, 422 (1958).
[18] P. Howard-Flanders, Nature **180**, 1191 (1957).
[19] T. H. Wood and A. M. Rosenberg, Biochim. et Biophys. Acta **25**, 78 (1957).
[20] A. M. Rosenberg, Ph.D. thesis, University of Pennsylvania (June, 1958).
[21] T. H. Wood and A. L. Taylor, Radiation Research **7**, 99 (1957).
[22] H. T. Meryman and W. T. Platt, Naval Medical Research Institute Project NM 000 018.01.08 (January 3, 1955).
[23] A. C. Stewart and J. A. Ghormley, Radiation Research **1**, 230 (1954).
[24] T. H. Wood, Ann. Rev. Nuclear Sci. **8**, 343 (1958).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

35

Effects of Radiations on Populations of Cells and Multicellular Organisms

CORNELIUS A. TOBIAS

Donner Laboratory of Biophysics and Medical Physics, University of California, Berkeley 4, California*

THE animal body is an assembly of cells. In considering the effects of radiation on such an assembly, the relationships between the individual cells and their assembly must be defined and, moreover, performance criteria must be found by which to characterize both the radiation effects on individual cells and the relation of these effects to the performance of the cell population.

There are three important principles which characterize the biological effects of penetrating radiations:

A. Radiation effects at the subcellular level are of a statistical nature. A single ionizing particle may lead to a decisive single radiobiological event in one of the many DNA nucleoprotein molecules or in another essential molecule. As a result, many different radiobiological events occur in the cells of an irradiated population.

B. In establishing the effect of irradiation on a cell population, the measured results are determined by the defined criterion for an experimental test. Because of the replication of genetic alterations in mitotic cell division, the criterion used for measurement frequently involves the selection of certain more-fit cells and the rejection of others, leading to end results different from what might be expected from statistical averaging of the entire population.

C. Radiobiology is an important basic tool because its fundamental interactions occur at the atomic-molecular level in essential sites of the cell, sometimes in locations not accessible to the ordinary tools of physics and chemistry. These interactions are then amplified to a level involving the entire cell—i.e., by a factor of about 10^{12}. Similarly, radiobiologic events that happened to a single cell or to a few cells may be amplified to the entire animal or human organism affecting the fate of as many as some 10^{14} cells.

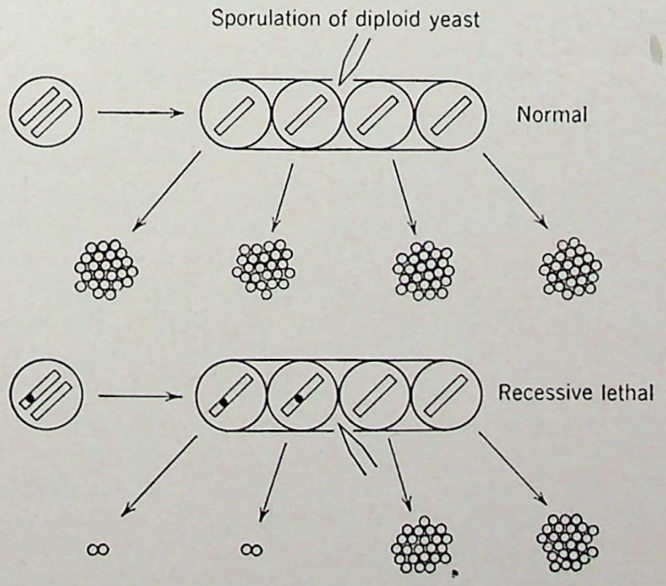


FIG. 1. Sporulation of normal and recessive lethal diploid yeast cells.

CELL-DIVISION DELAY AND INHIBITION

Zirkle (p. 269) and Wood (p. 282) have admirably outlined the major causes of cell-division delay and inhibition in irradiated yeast cells. In order to consider population effects in detail, the discussion of the same cells is continued with the understanding that in other species many additional interesting and diverse radiobiological phenomena occur. One cannot describe in a single chapter the wealth of phenomena that have been observed.†

The following classes of genetic damage are of interest:

A. Dominant lethal damage. This leads to inhibition of colony formation in one, or in a few, cell divisions.[1]

B. Recessive lethal damage[2,3] can kill the haploid cell, but will allow survival of the diploid. In some forms of this damage, the genetic material appears to be rearranged so that progeny of such cells, for many generations, divide slower than normal cells. When the diploid which bears recessive lethal damage is sporulated, it produces two dead spores for each recessive damage. The presence of recessive lethal damage in a diploid "softens" the cell for reirradiation, already suggested by Latarjet and Ephrussi,[4] and the survival of such cells indicates greater sensitivity than the normal diploid, as shown by Tobias and Stepka[5] and by Beam.[6] Recessive lethal radiation damage is illustrated schematically in Figs. 1 and 2.

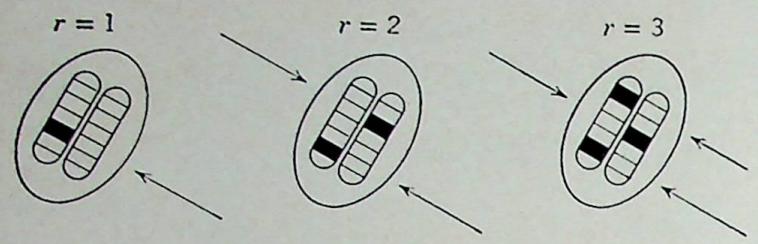
EXTENT OF DAMAGE IN NUCLEAR MATERIAL

The existence of radiation-induced auxotropic mutants of yeast cells in haploid mating types, requiring specific nutrients, makes it feasible to carry out genetic analysis of the damage induced. By mating individual cells under the microscope, one can breed diploid cells

* Work at the Donner Laboratory reported in this paper was performed under the auspices of the U. S. Atomic Energy Commission.

† A good idea of the complexity of radiobiological research may be obtained by consulting the journal, Radiation Research, and the book, *Proceedings of the International Congress of Radiation Research, Vermont, 1958*, to be published by the Academic Press, Inc.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA



$$N = N_0(2e^{-\alpha D} - e^{-2\alpha D})^{n-r} e^{-r\alpha D}$$

N = survivors
D = dose
n = number of compartments
r = number of defects
α = constant

FIG. 2. Recessive model of radiation survival of diploid yeast cell with defects.

with recessive requirements for specific nutrients. Irradiation of such cells often makes the requirement dominant due to production of damage in allelic sites. In this fashion, it is found that damage often extends to regions greater than a single biochemical locus; moreover, the probability of damage occurring from site to site is very different. The situation reminds one of chromosome breaks, which are seen to be produced copiously in other organisms. The damage extends most of the time to chromosome parts which include more than one gene. In relatively rare events, the damage extends to a single genetic locus only or to some part of it.[7]

POSTIRRADIATION CHANGES IN NUCLEAR MATERIAL

There are gene mutations that are extremely stable and persist through very many generations. It is believed generally that these mutations persist, and do not recover, and that their production varies linearly with dose. Except perhaps in some mutations of haploid cells, this statement is not exactly true, when one regards all kinds of nuclear damage from radiation.‡ In cells of higher ploidy, the presence of genetic material in two or more sets allows postirradiation rearrangements in the course of several subsequent cell divisions. The fate of a typical diploid survivor in the postirradiation period is shown in Fig. 3, as reported by Tobias.[7] Here, the irradiated cell was placed on nutrient agar under a microscope and observed continuously. As cell divisions occurred, the mother and daughter cells were separated. It is seen from the graph that, in the course of seven subsequent cell divisions, several cells were produced that failed to divide again. The radiated cells also exhibited long cell-division delays. A typical chart showing how a normal cell gives rise to progeny is given also. The process of generation of cells lacking viability can go on for many generations so that some colonies arise where cell replication is below optimum value. For instance, by preirradiation it is easy to obtain colonies which produce one dead cell for every four normal ones.

One of the suggested mechanisms for such a postirradiation recovery effect is shown in Fig. 4. As the cell with recessive radiation damage in its nucleus divides, occasionally "chromosome crossover" occurs; i.e., segments of the chromosome exchange places with each other. The two progeny are not identical in chromosome content. In the instance shown, one of them has homozygous lethals, whereas the other will survive, having gotten rid of the damaged part. A method of experimentally demonstrating chromosome crossover uses recessive genetic markers. In the example shown, an adenineless locus was carried in the heterozygous state. When homozygous, the cells needing adenine become pink; otherwise they are ivory colored. Thus, a clone established from one irradiated diploid cell may appear color segmented, with approximately $\frac{1}{2}$, $\frac{1}{4}$, or $\frac{1}{8}$ of the cells colored, depending upon whether crossover occurred in the first, second, or third postirradiation cell division. This sort of delayed phenotypic appearance of induced chromosome changes has been known since the work with bacteria by Witkin, Demerec, and Latarjet.[8–10] Taking advantage of the detailed genetic knowledge available in yeast cells, James demonstrated ultraviolet-induced homozygosis[11] and Mortimer demonstrated the x-ray-induced phenomenon.[12]

Appearance of homozygosis, following damage, of recessive genetic traits is a defense mechanism, but it also serves to make recessive lethals existing in the population homozygous.

There are other nuclear changes suspected of occurring in the postirradiation period. In some large plant and animal cells, these are the consequences of chromosome breaks and their rejoinings. The result may be wholly- or partially-increased ploidy, euploidy, or aneuploidy.

CHANGES IN RADIORESISTANCE

In a mixed population of cells, various stages of cell division are present. It has been shown in haploid yeast cells that, in the early phase of budding, the cells are more than ten times more resistant than in inter-

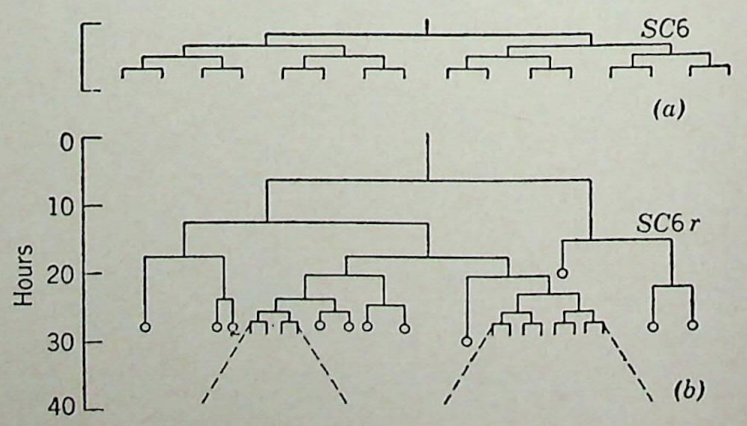


FIG. 3. Postirradiation divisions of diploid cell. (a) Nonirradiated. (b) Irradiated. Open circles indicate cells which have stopped dividing.

‡ *Note added in proof.*—Most of the evidence for linearity comes from radiation studies on the sperm of the fruit fly. Very recently, Russell *et al.* [Science 128, 1546 (1958)] have furnished experimental proof of the dose-rate dependence of some mutations by irradiating male and female mice with x-rays and γ-rays.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

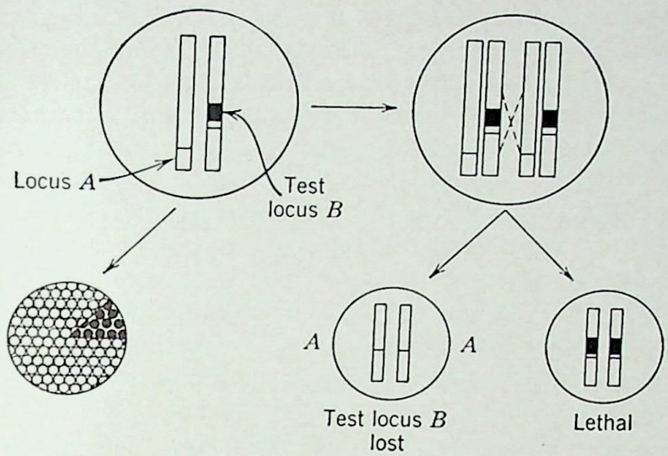


FIG. 4. Mechanism of postirradiation effect—chromosome crossover. If crossover occurs in delayed fashion, one may observe the appearance of segmented clones of cells, indicated on the left.

phase.[6,13] Moreover, kinetics of the sensitivity to radiation are different; while a single ionizing event in the proper location is sufficient for inhibition during interphase when most new DNA synthesis occurs, dividing cells require some 100 such events. Accepting DNA nucleoprotein as the site of action,§ this phenomenon is a clear indication that physical and perhaps chemical states of this substance must change radically in the course of cell division. DNA molecules undergoing synthesis are highly sensitive; existing DNA molecules exhibit decreased sensitivity. It would appear that, during mitotic division of yeast, the building blocks of the new chromosome are present in great multiplicity, after which they condense to make just two new nuclei.

RECOVERY FROM MUTATIONAL DEFICIENCIES

Point mutations induced by ultraviolet- or x-rays generally have a low but definite spontaneous reversion rate in the haploid cell. It appears that within the same biochemical locus many kinds of damage may be produced and that each of these has a characteristic reversion rate. Since many kinds of mutations may be prepared by radiation at the same locus, these eventually may serve to help in the understanding of the chemical nature of the genes and their ability to transmit information to other cell constituents.‖

When a diploid cell is made up of two haploids with heteroallelic auxotropic deficiency at the same locus, the cells cannot grow into colonies in the absence of the required nutrient. However, as Roman and Jacob have shown, ultraviolet irradiation of such a cell can bring about a striking reversion to independence of the same nutrient.[14] This occurs at unusually low doses and may be regarded as a beneficial effect of radiation, since it induces repair. The mechanism is not known. It is as though the irradiation helped the cell to make faultless replicas from the imperfect parts of the two allelic genes, a mechanism sometimes called the *copy-choice* method of replication.

CYTOPLASMIC EFFECTS

While the cell nucleus undergoes injury, cytoplasmic deficiencies also take place. There is every reason to believe that RNA enzyme molecules and proteins are equally as vulnerable as DNA. Even if such injury occurs, however, the situation for the cell usually is not assumed to be too serious, since there are many identical RNA molecules and since it is assumed that, with the aid of DNA, they may be resynthesized; on the other hand, the loss of function of an essential DNA molecule may prove fatal to the cell.

There are some autonomous cytoplasmic particles in various cells. According to the work of Ephrussi,[15] these particles in yeast are known to be able to duplicate themselves and the rate of their duplication may occur somewhat independently from that of DNA.[16] When the granules carrying the cytochromes are lost from the cells, they cannot be resynthesized. Radiation produces these small-colony mutants, with ultraviolet- being much more efficient than x-rays (as first shown by Raut[17]).

Almost any other cytoplasmic variable tested has been shown to be effected by radiation, though generally requiring higher doses than do the nuclear effects. Examples consist of disturbances in ionic balance, release of ATP and other substances from the cells, changes in activity of various enzymes, etc. These effects in yeast cells are reviewed by Rothstein.[18]

EXTRACELLULAR EFFECTS

Extracellular effects also are of importance in studying cell populations. These include the following factors:

A. Radiation effects on the nutrient medium.

B. Decrease in the autocatalytic action of cells which may synthesize important substances for other cells.

C. Release of growth-promoting and radiation-protective substances by the irradiated cells into the medium.

Interestingly enough, the last item of these seems to be of greatest importance. It first was detected in yeast cells under ultraviolet light by Loofbourow some 20 years ago, and more recently has been demonstrated for x-rays by Gunter and Kohn.[19] In mammalian tissue culture, Puck and Marcus[20] have shown such effects, while Révész[21] has worked on the problem in detail for ascites cells.

CYTOLOGICAL EFFECTS OF PENETRATING RADIATIONS IN TERMS OF INFORMATION THEORY

Radiations generally decrease the information content of the genetic apparatus. They also increase the

§ Although one generally assumes that chromosome damage occurs as a result of a single ionizing particle, it has never been proved that this primary interaction is with the DNA molecule. Release of an enzyme capable of acting on DNA may have the same effect, and such an assumption indeed would explain the chain reaction which results in the breaking of whole chromosomes.

‖ See chapters by Levinthal (p. 249) and by Lennox (p. 242).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

noise of the information channels—namely, extragenic components—which are instrumental in transmitting instructions for cellular functions, particularly duplication. This necessitates a lag in cell division which has a definite limit, however, and, if complete information is not transmitted in due time, the new cell is lost. Systems with redundancy in the genetic apparatus can transmit the information faster and more reliably.

Radiation damage to the coding system not only consists of the knocking out of individual symbols, but also consists of deletions of extensive parts of the code or of their rearrangement in spatial order.

The cell has mechanisms to repair the damaged genetic code to some extent. Given time, it can develop a method of more-reliable information transmission even in the presence of continuous radiation which is producing new damage at a steady rate. Recovery may occur by increasing the channel capacity and, in some instances, by correcting the code or increasing its redundance. In the course of such events, groups of cells with maximum equivocation or greatest information-transfer efficiency receive preference and may take over the colony.

POPULATIONS OF CELLS IN THE ANIMAL BODY

Radiation effects in the animal body are consequences not only of irradiation of its cells, but also of changes in functions of its organs and in their interactions with each other. For purposes of this discussion, an organ may be regarded as a heterogenous population of cells in a continuous state of activity—growth and anabolism for some cells, decay and catabolism for others. The cells are bathed in body fluids, which bring fresh nutrients and eliminate the products of cell metabolism and, at times, some of the cells themselves.

YEAST-CELL POPULATIONS IN THE STEADY STATE

To approximate conditions in an organ, Welch in our laboratory[22] worked with a population of diploid yeast cells, propagated continuously in a device resembling the chemostat of Novick and Szilard.[23] A schematic view of this device is seen in Fig. 5. When a nutrient medium flows into this system at a given rate and when a medium with cells is eliminated from it at the same rate, steady-state populations of cells are established with a constant rate of cell division. Welch has irradiated such cultures continuously with x-rays for more than one hundred consecutive cell divisions. As the radiation delayed cell division, the flow rate of the chemostat was adjusted so that a new steady state of cell population was established at the same cell density as in the controls. In a population of this kind, radiation can cause killing by the dominant lethal mechanism. The recessive lethal mechanism can also kill, either by the change of heterozygosis into homozygosis, or by the accumulation of homozygous recessive lethals. If one assumes the view that no recovery of gene-radiation damage takes place, then it follows that in each generation the population can tolerate a very small fraction of the dose required for acute killing.

Actually, the population has maintained continuous proliferation up to the highest dose rate tested, about 6000 r/generation time. The main effects of radiation were the prolongation of the time required for cell division, and the generation of some cells incapable of further cell division.

From the number of viable and nonviable cells present, it is possible to evaluate the mean lifetime of cells in the population and the average number of daughter cells a mother cell is capable of producing before her ability to undergo mitotic division is lost.

Figure 6 indicates how the number of daughter cells from each mother cell decreased rapidly with increasing dose rate. Each control cell produced about 66 daughters; this number decreased to about seven at the highest dose rate. At the same time, however, the cells appeared to acquire greater radioresistance by a factor of two than they had had prior to irradiation. The increased radioresistance appeared to develop by adaptation, rather than by mutation. A few generation times after radiation was stopped, the cells regained their normal rate of cell division. Cells taken out of the steady-state culture and irradiated in conditions of starvation showed normal radioresistance.

Although a great deal more needs to be done with continuously proliferating cellular systems, Welch has demonstrated that diploid populations can live in the presence of surprisingly large radiation levels, and that part of the radiation-induced damage recovers in the course of successive cell divisions.

RADIATION EFFECTS ON MAMMALS AND MAN

It is impossible to do justice to the intricate and detailed research work that has been going on in radia-

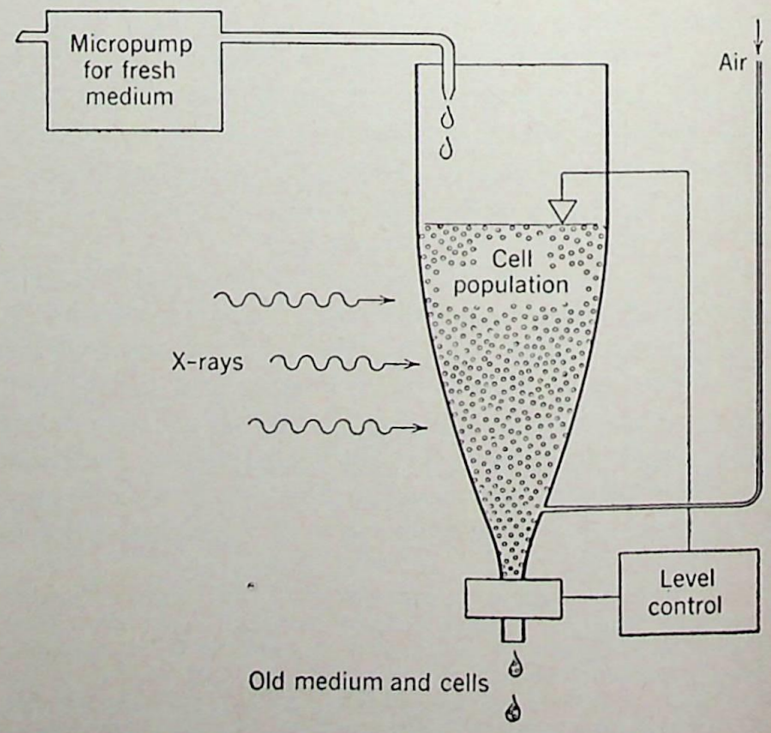


FIG. 5. Radiation of yeast-cell populations in chemostat.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

tion physiology during the past decades. The reader is referred to the excellent review articles and books on the subject.[24–26] It is important to note that bacteriological techniques of propagation recently have been applied successfully to human cells by Puck¶ and that, as a result, some information is already available on the radiosensitivity of various diploid and tetraploid human-cell strains. The mean lethal doses range from about 90–300 r, and the mechanisms of inhibition of colony formation are perhaps not too different from those for yeast cells. It is not known, however, what the radiation resistance of differentiated human cells in the body is, where it is possible that the very rapid humoral and metabolic exchange and the changes of radiosensitivity due to cell division might lead to an increase of radioresistance, as noted in the steady-state yeast-cell cultures.

RADIOSENSITIVE TISSUES

There is a 60-year-old law in radiobiology, originally proposed by Bergonnier and Tribondot; according to this law, the most radiosensitive tissues are those which have the highest mitotic index—i.e., the highest rate of cell division. One now understands that this law is only approximately true because radiation injury manifests itself mainly in the cell-division process. Radiated cells usually die when they attempt cell division; if their functions in body tissues do not require frequent cell division, the cells usually can perform ordinary metabolic and enzymatic functions for a long time post irradiation, since these functions are much less radiosensitive than cell division itself.

Correspondingly, in the acute phase, a single dose of radiation usually affects the most rapidly proliferating tissues; the bone marrow and its varied cell population and certain epithelial tissues are affected first. Early radiation deaths frequently are due to injury to the intestinal mucosa and to white-cell proliferation, while failure of the blood clotting mechanism and anemia make their appearance later. The neurons of the central nervous system of the adult are quite radioresistant, since almost no cell division occurs in them. In the developing mammalian embryo, however, as shown by Hicks,[27] nerve tissue is rapidly proliferating and has the greatest radiosensitivity.

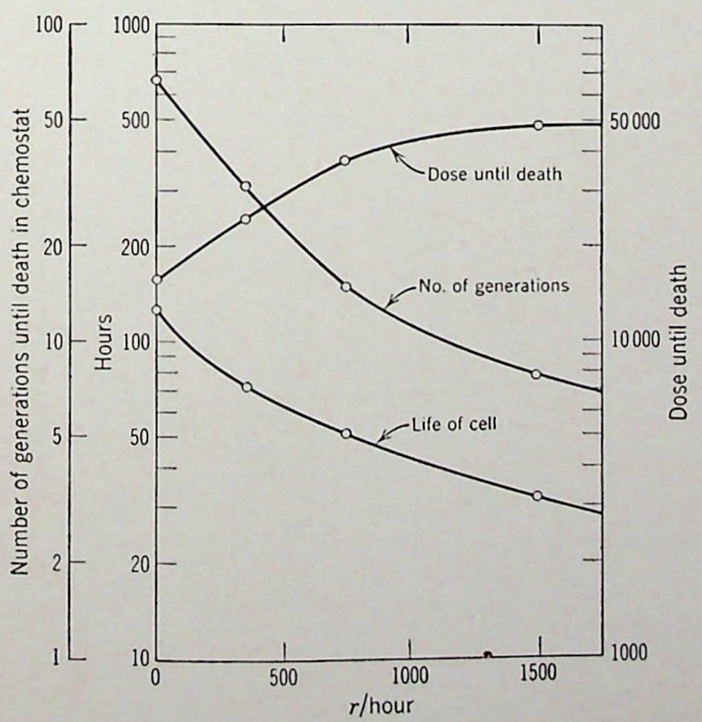


FIG. 6. Progeny decrease with dose rate.

¶ See chapter by Puck (p. 433) for detailed discussion.

INDIRECT RADIATION EFFECTS

There are very intricate relationships between various organs of the body, and it is not surprising that remote effects of radiations have been observed. For example, if an organ receives a dose of radiation, depressive effects due to this injury manifest themselves elsewhere in the body. Starting with the work of Hevesy, it was shown that a small depression of cell division and of DNA synthesis occurs everywhere in the body if an organ, such as the spleen, is irradiated.[28] To illustrate remote effects of radiation, one may mention work done at the Berkeley cyclotron in irradiating pituitary and various regions of brain tissue by high-energy deuterons.[29] Because of their small scatter, these particles are well suited to make small radiolesions in various parts of the body. It was found that a large localized dose to the rat pituitary (size 1×2 mm) will cause progressive development of the hypophysectomized state.[30] The growth of the animal, calcification of its bones, and functions of adrenals, thyroid, and gonads are all affected; development of the remote effects has different dose-response curves for each kind of effect.

An organ that is shielded from radiation confers a degree of protection against effects from whole-body radiation to the animal. Shielding of spleen and bone marrow is of greatest benefit, sometimes increasing radiation tolerance by about a factor of two. Jacobson and his associates[31] have shown that transplantation of bone marrow, spleen, or embryo homogenate to mice which have received a dose of x-rays increases radiation tolerance. The greatest part of this effect is owing to transplantation of healthy, unirradiated bone-marrow cells, which prosper on the tissue bed of radiation-killed bone marrow. Apparently, one may allow death of all cells of the marrow by whole-body x-irradiation, and the animal may be resurrected by the administration of a sufficient quantity of serologically compatible healthy cells. This technique has given rise to valiant attempts now being made in various medical installations to save terminal leukemia patients. Patients receive a lethal dose of whole-body radiation, followed by a bone-marrow transplant obtained from healthy relatives. The radiation is supposed to kill all bone-marrow cells, including the leukemic cells. If the transplanted marrow saves the patient from the radiation effects, it

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

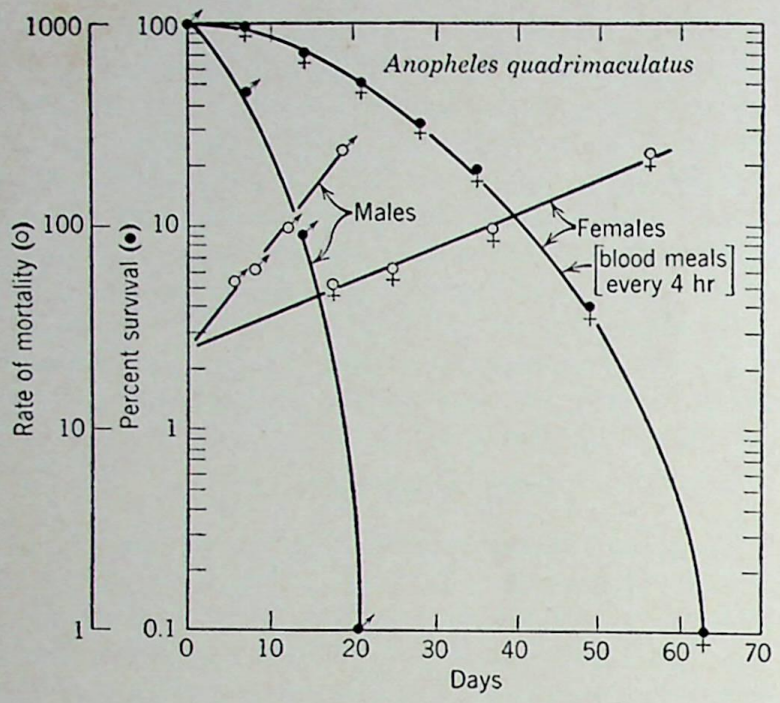


FIG. 7. Population-survival curve—mosquitoes [from W. E. Kershaw, T. A. Chalmers, and M. M. J. Lavoipierre, Ann. Trop. Med. Parasitol. 48, 442 (1954)].

is hoped that leukemia will not recur. Because of immunological reasons, much research is needed before bone-marrow transplants are completely safe for humans.[32]

LONGEVITY

In normal populations of humans, animals, invertebrates, and even unicellular organisms, there appear to be identical laws predicting the rate of death of the population. About a hundred years ago, Gompertz worked out details of this function, and at the present time Jones *et al.*[33] have applied it in detail to many problems of population survival.

The relative rate of death of a population of N individuals as function of time t is

$$\frac{d^2 \log N}{dt^2} = \alpha t,$$

where α is a characteristic constant of the population. To illustrate this principle, a normal survival curve of mosquitoes is reproduced in Fig. 7.[34] The relative rate of death is an exponentially increasing function. In adult humans, the death rate is described by a similar exponentially increasing function, with the rate doubling itself about every eight years. According to the theory by Jones, a single dose of radiation will modify the relative rate of death by shifting it upwards, parallel to itself. Continuous irradiation is supposed to increase the rate constant, α. Although many longevity estimations were made on the basis of this theory which can be used for daily permissible-dose determinations, rigorous experimental proof for radiation effects is not as yet available. Several other theories have been proposed.[35]

CARCINOGENESIS

It is of great interest to know something of radiation carcinogenesis. Cancer cells represent a population which is abnormal in the sense that (a) it has become independent of some of the growth-controlling factors characteristic of the tissue from which its cells originate; (b) it is frequently able to metastasize; and (c) it may have metabolic products which lead to death of the host. The rate of incidence of "spontaneous" cancer, chemical-, virus-, or radiation-induced neoplasms, usually, at least in some definite time interval, follows a law similar to that propounded for the rate of death of populations: The rate of onset increases as an exponential function of time. As an illustration, the rate of onset of bone tumors resulting from radioactive Sr^{89} incorporated in bone is shown in Fig. 8; this work was done at the Argonne National Laboratory.[36**] The rate constant of tumor incidence increases with increased radiation dose and is a more important index of carcinogenesis than is the over-all increase of tumors in a lifetime, which, in addition to the cancer-inducing factor, depends upon factors determining longevity. In practice, such an exponential rate curve means a "time lag" of varying length before the demonstrable onset of neoplastic growth.

The detailed biochemical steps that lead to cancer are not known, but it is useful to draw an analogy between diploid cells of animal tissue that have received a dose of radiation and diploid yeast cells, discussed earlier in this paper. An individual, sublethally irradiated yeast cell frequently develops an abnormal colony distinguished by its slow rate of growth in the process of "recovery." Some of the cells regain their full ability to proliferate by processes outlined in the section in which the recovery from radiation damage was dis-

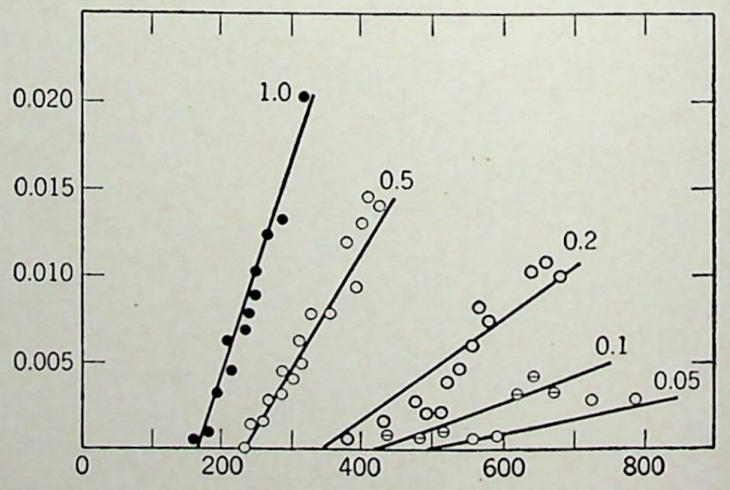


FIG. 8. Daily probability of bone-tumor development in the mouse following monthly injections of Sr^{89} (after Brues[36]). Abscissas: time after first injection (days). Ordinates: daily probability of bone-tumor development. Curves show effect of monthly doses of Sr^{89}, of 1.0, 0.5, 0.2, 0.1, and 0.05 μC/g.

** *Note added in proof.*—A. Brues and M. Finkel have published recent interpretations of their experimental work in carcinogenesis [Science **128**, 637, 693 (1958)]. Examples of other views on this subject are found in the work of E. B. Lewis [Science **125**, 965 (1957)] and in references 36, 40, and 42.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

cussed. This process often takes many cell generations, and in the course of it a number of changes can occur that will modify the genetic constitution of the cell. Changes resembling carcinogenesis in yeast cells also have been discussed by Lacassagne *et al.*,[37] Maisin *et al.*,[38] and Warburg.[39]

Most human cells normally are under hormonal and possibly neural control of the homeostatic apparatus and, as a result, they usually are in a state of proliferation which is below their maximum proliferative capacity. As an illustration, consider the human breast. This organ develops under elaborate hormonal influence, responding to secondary sex hormones elaborated by the ovaries and adrenals and to lactogenic hormone. In addition, the cells of this gland are influenced by growth hormone and posterior pituitary hormones. The state of this tissue and its metabolic activities are conveyed in some way to the hypothalamus and the pituitary, and as part of homeostatic control the gland may receive increased hormonal stimulation when its performance is below par.

There is increasing evidence available to show that hormonal balance is an important factor in radiation carcinogenesis. The intricate pattern of the hormonal factors has been studied for more than 80 years; here a few recent experiments only are noted, with mammary cancer as a restrictive example. Lacassagne has shown clearly that mammary cancer in mice may be induced by estrogen in females, as well as in castrated males. More recently, it appears that the joint application of estrogen and lactogenic hormone accelerates the onset of cancer. Bond *et al.*[40] found that a sublethal dose of whole-body irradiation in a strain of rats induced mammary cancer. However, the frequency of mammary cancer was decreased greatly if the ovaries of the animals were removed, thus eliminating most of the endogenous, secondary sex hormones. It is known that, following subtotal I^{131} thyroidectomy in mice, thyroid tumors as well as pituitary tumors result. Furth *et al.*[41] found that some rats recovering from whole-body radiation developed pituitary tumors, apparently in response to hormonal demands from endocrine target organs. When such tumors are transplanted into other rats of the same inbred strain, hormonal stimulation from the graft includes adrenal, thyroid, and mammary tumors in the new hosts. Apparently, whenever some tissue in the hypothalamic-pituitary-endocrine organ system develops sublethal tissue damage involving chromosome derangements, so that the tissue is unable to perform its functions satisfactorily, there is strong hormonal and/or nervous stimulation of these tissues causing increased rate of proliferation. Under such conditions, genetic rearrangements may lead to neoplastic tissues. An interesting illustration is furnished by localized deuteron irradiation of the rat pituitary at Berkeley.[29] Here 945 rep to the rat pituitary led within one year to the development in every irradiated animal of pituitary tumors, with 3.5 tumors per gland. Higher doses led to lessened incidences of tumors, and when most pituitary cells were killed, none of the animals developed pituitary tumors.

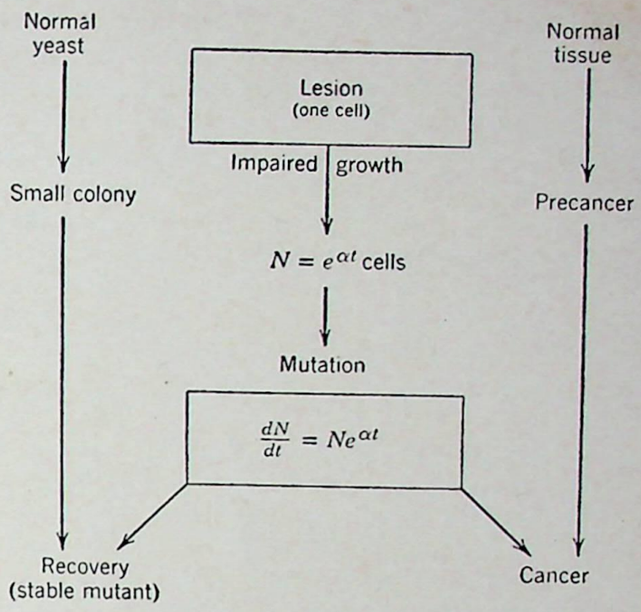


FIG. 9. Radiation carcinogenesis.

There are many other instances of hormonal influence over carcinogenesis. Perhaps the most complete investigations of this kind were made by Kaplan and his group.[42]

It would appear that a better understanding of homeostatic control mechanisms may lead to an understanding of the etiology of cancer and possibly to some methods of prevention or delay of its onset. In the meanwhile, it has been apparent that cancer cells that do not become completely undifferentiated in the carcinogenic process still retain some measure of dependence on hormonal control. Work is in progress in many countries to hypophysectomize patients with advanced and progressing metastatic carcinoma, in the hope that the cancer cells will stop growing and the tumors will regress.

The proton- and α-particle beams of the Berkeley 184-in cyclotron have proved to be promising tools for research on hypothalamic functions and for therapeutic investigations involving radiation hypophysectomy. The techniques and initial results have been described.[43] While hypophysectomy is not the final answer to cancer therapy, there are some patients who appear to have benefited from the pituitary irradiation. Detailed biochemical study of induced tumor regressions may bring additional clues with respect to the nature of the hormones which control normal or abnormal proliferation.

In conclusion, radiobiology has been useful as an aid to understanding not only the complex physiological changes following irradiation exposure, but also, the mechanism of carcinogenesis. Basic research on molecules and cells has helped in the understanding of some phenomena of cell division and of the nature of some

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

genetic transformations. In Fig. 9, the present status of understanding of radiation carcinogenesis is shown in simplified fashion. It is apparent from the figure that there is a qualitative similarity between postirradiation events in diploid yeast cells and in somatic tissue cells. Many of the diploid yeast survivors develop small-colony mutants, some of them cytochrome deficient, that proliferate slowly and revert to rapid cell division many generations later. This process may be analogous to the slow development of benign precancerous lesions and the subsequent somatic mutation to cancerous cells. The time rate of occurrence of both sets of phenomena is quite similar. It is clear, then, that more work is needed to ascertain the genetic and biochemical features of delayed recovery in irradiated unicellular organisms and changes in the aerobic-anaerobic metabolism which, according to Warburg,[39] are known to occur in carcinogenesis. Perhaps such work may furnish some day direct clues to the detailed etiology and inhibition of cancer.

BIBLIOGRAPHY

[1] R. K. Mortimer, Radiation Research **2**, 361 (1955).

[2] R. E. Zirkle and C. A. Tobias, Arch. Biochem. Biophys. **47**, 282 (1953).

[3] R. K. Mortimer and C. A. Tobias, Science **118**, 517 (1953).

[4] R. Latarjet and B. Ephrussi, Compt. rend. **229**, 306 (1949).

[5] C. A. Tobias and B. Stepka, "Mutations to increased radiosensitivity in yeast cells" [University of California Radiation Laboratory Report (UCRL-1922), 1952], p. 40.

[6] C. A. Beam, "Influence of metabolism on some radiation effects," in *Proceedings of International Congress of Radiation Research, Vermont, 1958* (to be published).

[7] C. A. Tobias, R. K. Mortimer, R. L. Gunther, and G. P. Welch, "The action of penetrating radiations on yeast cells" [University of California Radiation Laboratory Report (UCRL-8268), 1958]. To be published in *Proceedings of Second United Nations International Conference on the Peaceful Uses of Atomic Energy, Geneva, 1958.*

[8] E. M. Witkin, Proc. Natl. Acad. Sci. U. S. **32**, 59 (1946).

[9] M. Demerec and R. Latarjet, Cold Spring Harbor Symposia Quant. Biol. **11**, 38 (1946).

[10] M. Demerec, Proc. Natl. Acad. Sci. U. S. **32**, 36 (1946).

[11] A. P. James, Genetics **40**, 204 (1955).

[12] R. K. Mortimer, "Radiobiological studies on yeast (discussion)," in *Proceedings of International Congress of Radiation Research, Vermont, 1958* (to be published).

[13] C. A. Beam, R. K. Mortimer, R. G. Wolfe, and C. A. Tobias, "The relation of radio-resistance to budding in saccharomyces cerevisiae" [University of California Radiation Laboratory Report (UCRL-2345), 1953]; Arch. Biochem. Biophys. **49**, 110 (1954).

[14] H. Roman and F. Jacob, Compt. rend. **245**, 1032 (1957).

[15] B. Ephrussi, *Nucleo-Cytoplasmic Relations in Micro-Organisms, Their Bearing on Cell Heredity and Differentiation* (Clarendon Press, Oxford, 1953).

[16] F. Sherman, "A study of the effects of elevated temperatures on the growth and inheritance of Saccharomyces cerevisiae," thesis, University of California, Berkeley (1958) [University of California Radiation Laboratory Report (UCRL-8573), 1958].

[17] C. Raut and W. L. Simpson, Arch. Biochem. Biophys. **57**, 218 (1955).

[18] A. Rothstein, "Biochemical and physiological changes (in yeast cells)," in *Proceedings of International Congress of Radiation Research, Vermont, 1958* (to be published).

[19] S. E. Gunter and H. I. Kohn, J. Bacteriol. **72**, 422 (1956).

[20] T. T. Puck and P. I. Marcus, J. Exptl. Med. **103**, 653 (1956).

[21] L. Révész, J. U. S. Natl. Cancer Institute **20**, 1157 (1958).

[22] G. P. Welch, "Effects of chronic exposure to x-rays on a steady rate population (of Saccharomyces cerevisiae)." Thesis, University of California, Berkeley, (1957); University of California Radiation Laboratory Report (UCRL-3763), 1957.

[23] A. Novick and L. Szilard, Science **112**, 715 (1950).

[24] A. Hollaender, editor, *Radiation Biology* (McGraw-Hill Book Company, Inc., New York, 1954–1956), Vols. I–III.

[25] J. J. Nickson, editor, *Symposium on Radiobiology, Oberlin College, 1950. The Basic Aspects of Radiation Effects on Living Systems* (John Wiley and Sons, Inc., New York, 1952).

[26] Z. M. Bacq and P. Alexander, *Fundamentals of Radiobiology* (Academic Press, Inc., New York, 1955).

[27] S. P. Hicks, Am. J. Roentgenol. Radium Therapy Nuclear Med. **69** (2), 272 (1953).

[28] L. S. Kelly and H. B. Jones, Am. J. Physiol. **172**, 575 (1953).

[29] J. L. Born, A. O. Anderson, H. O. Anger, A. C. Birge, P. Blanquet, T. Brustad, R. A. Carlson, D. C. Van Dyke, D. J. Fluke, J. Garcia, J. P. Henry, R. M. Knisely, J. H. Lawrence, C. W. Riggs, B. Thorell, C. A. Tobias, P. Toch, and G. P. Welch, "Biological and medical studies with high energy particle accelerators" [University of California Radiation Laboratory Report (UCRL-8242), 1958]; in *Proceedings of the Second United Nations International Conference on the Peaceful Uses of Atomic Energy, Geneva, 1958* (to be published).

[30] D. C. Van Dyke, M. E. Simpson, A. A. Koneff, and C. A. Tobias, "Long term effects of deuteron irradiation of the rat pituitary" [University of California Radiation Laboratory Report (UCRL-8238), 1958]; Endocrinology (to be published).

[31] L. O. Jacobson in *Radiation Biology*, A. Hollaender, editor (McGraw-Hill Book Company, Inc., New York, 1954), Vol. I, p. 1029.

[32] A. Hollaender, C. C. Congdon, D. G. Doherty, R. Shapira, and A. C. Upton, "New developments in radiation protection and recovery," Oak Ridge National Laboratory Report in *Proceedings of the Second United Nations International Conference on the Peaceful Uses of Atomic Energy, Geneva, 1958* (to be published).

[33] H. B. Jones in *Advances in Biological and Medical Physics* (Academic Press, Inc., New York, 1956), Vol. IV, p. 281.

[34] W. E. Kershaw, T. A. Chalmers, and M. M. J. Lavoipierre, Ann. Trop. Med. Parasitol. **48**, 442 (1954).

[35] N. Berlin and F. L. DiMaggio, "A Survey of Theories and Experiments on the Shortening of Life Span by Ionizing Radiations" [Armed Forces Special Weapons Project Report (AFSWP-608), 1956].

[36] A. M. Brues and M. P. Finkel, quoted in R. H. Mole, Brit. Med. Bull. **14**, 184 (1958).

[37] A. Lacassagne, M. Schoen, and P. Beraud, Ann. Fermentations **5**, 129 (1939).

[38] J. Maisin, G. Lambert, and E. Van Duyze, Rev. belge Pathol. et méd. exptl. **21** (1), 52 (1951).

[39] O. Warburg, Science **123**, 309 (1956).

[40] V. P. Bond, E. P. Cronkite, C. J. Shellabarger, S. W. Lippincott, R. A. Conard, and J. Furth, "Mechanisms of radiation-induced mammary gland neoplasia related to dose dependence and hormonal status in rats," in *Proceedings of Second United Nations International Conference on the Peaceful Uses of Atomic Energy, Geneva, 1958* (to be published).

[41] J. Furth, E. I. Hirsch, H. J. Curtis, E. L. Gadsden, and R. F. Buffett, "Pathogenesis and character of radiation-induced pituitary tumors," presented before the Radiation Research Society, Rochester, New York, 1957; abstract: Radiation Research **7** (3), 317 (1957).

[42] H. S. Kaplan, C. S. Nagareda, and M. B. Brown in *Proceedings of the Laurentian Hormone Conference*, G. Pincus, editor (Academic Press, Inc., New York, 1954), Vol. X, p. 293.

[43] C. A. Tobias, J. H. Lawrence, J. L. Born, R. K. McCombs, J. E. Roberts, H. O. Anger, B. V. A. Low-Beer, and C. B. Huggins, Cancer Research **18**, 121 (1958).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

REVIEWS OF MODERN PHYSICS VOLUME 31, NUMBER 2 APRIL, 1959

36

Fine Structure of Cell Nucleus, Chromosomes, Nucleoli, and Membrane

H. STANLEY BENNETT

Department of Anatomy, University of Washington School of Medicine, Seattle 5, Washington

THE most important concepts of the structure of the nucleus are based on information drawn from three rather separate approaches which are now converging. These approaches had their origins in the nineteenth century. The first of these—the optical approach—had its origin with Robert Brown's first recognition of the nucleus by the use of the light microscope in 1833. This distinguished botanist was also the discoverer of what is now called Brownian movement. The amplification of his original observation, as carried out with various optical tools, forms an exceedingly important part of present knowledge of the nucleus.

The second stream of knowledge is that of genetics. It can be regarded as beginning with Mendel's observation that inherited characteristics are transmitted from generation to generation in ratios of simple whole numbers. During the early part of the present century, derivatives of this discovery, correlated with optical observations on the nucleus, led to the field of cytogenetics. Subsequently, the information derived from genetic sources has played an essential part in concepts of nuclear structure. Indeed, the genetic data provide the only evidence regarding certain important structural features of the nucleus.

One can trace the third stream, the chemical one, to the work of Miescher, who published an important compendium in 1897 on the chemical characteristics of extracts derived from pus. Lamentably, this raw material was abundantly available in those days, before aseptic surgery had become popular. From this convenient product of human suffering, Miescher extracted substances which he named nucleic acids.

One can ask if the nucleus is an essential part of the cell. One can recall the red cell of the mammal, which is active for about 120 days without a nucleus. But how do other cells fare if the nucleus is removed? This has been investigated in a relatively small number of forms. It is possible to remove the nucleus from an amoeba by surgical means, as Mazia, for example, has done. It is possible also to amputate and study non-nucleated fragments of certain large algae, such as *Acetabularia*. One can summarize much by saying that a non-nucleated cell or cell fragment can survive for a certain period. An amoeba can eat, it can move about after enucleation, but in due time its synthetic capacities appear to degenerate, and, like the red cell, it perishes without reproducing. The red cell is perhaps the most successful of cells in functioning for a long period without a nucleus.

It is possible to isolate nuclei and to separate them from cytoplasmic components. If such isolated nuclei are studied biochemically, their metabolic behavior can be followed. The most elegant work along these lines has been carried out by Mirsky and Allfrey, based on earlier work by Dounce. The nuclei contain protein, DNA, RNA, and small amounts of lipid. Allfrey and Mirsky have found that isolated nuclei are capable of carrying out amino-acid incorporation into proteins and that certain other enzymatic activities are associated with the nucleus.

If one examines the unstained nucleus with a light microscope, either in the living cell or after fixation with ordinary reagents, one sees regions of varying density. This inhomogeneity of structure becomes even more striking following staining, since the nuclear components vary in their affinity for dyes. The components which bind the dyes have been called chromatin. As a rule, the term "chromatin" is used primarily to designate strongly-staining materials in the nucleus of the cell. It is a general term without any specific chemical significance. The term "nuclear chromatin" is applied to two distinct structures within nuclei. One comprises the chromosomes themselves; the second consists of nucleoli which are accumulations containing considerable quantities of RNA. Many of the staining reactions of chromosomes and of nucleoli are very similar or identical, but there are a few methods for distinguishing them from each other. These depend on chemical differences between DNA on the one hand, which dominates many of the properties of chromosomes, and RNA on the other, which has an important role in the nucleoli.

The appearance of nuclear chromatin can change greatly under different functional states of the cell. As the cell approaches a mitotic division, the chromatin material rearranges itself. At that time, one easily can see well-defined bodies, the chromosomes, so called because they can be stained so as to assume very vivid colors. It is well established that the chromosomes are present in the intermitotic nucleus, where they usually occur in a rather tenuous form which makes it difficult to recognize them morphologically. Yet chromosomal activity can be detected by a number of different means. The genetic method is perhaps the most powerful.

In certain specialized cases, the chromosomes may be present in a form quite different from the dispersed state which they assume in the normal intermitotic cell. For example, in the salivary gland cells of the fruit fly, *Drosophila*, the intermitotic chromosomes occur as dis-

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

crete bodies known as "giant" chromosomes, which may be over 200 μ long. Under the light microscope, they show a longitudinal fibrillar structure and a distinct series of aperiodic crossbands. These crossbands are unique in the sense that they exist in distinct and recognizable patterns in each chromosome. These patterns are important for correlating the genetic influence of the chromosomes with its structure.

Consider now the question: How can one exploit the properties of the various chemical components of the nucleus in such a way as to gain information about their disposition, structure, interactions, and relationships? Both acid and basic dyes are bound by many of the components of the nucleus. From this, it is inferred that nucleoproteins behave as a mixture of anionic and cationic ion-exchange resins. These properties can be attributed to their negatively charged phosphate groups and positively charged protein groups. One can exploit these characteristics by setting up a suitable competitive system which introduces colored cationic molecules to bind to the nucleic-acid phosphate groups, replacing positively charged protein groups. This procedure tags phosphate groups of the nucleic acids with colored tracers or indicators. The capacity of anionic polymer groups, such as nucleic-acid phosphate groups, to bind basic dyes is spoken of as the property of "basophilia." Thus, nucleochromatin is often spoken of as being strongly basophilic. In this way, the nucleic acid can be tagged rather nonspecifically. Analogously, one can use anionic dyes such as azosulfonic acids to bind to polymers containing basic groups in tissue components. Such polymers are found in the nucleus in the form of the basic proteins which are associated with the nucleic acids to form nucleoproteins.

Most of these dye-binding methods do not distinguish between RNA and DNA. But, from the analytical data on isolated nuclei, there is reason to believe that both are present in the nucleus. If one turns to another property of nucleic acids—namely, the absorption spectrum in the ultraviolet of the purine and pyrimidine residues—it can be shown that the same structures which take up basic dyes also absorb strongly in the ultraviolet. This is consistent with the view that the nucleoprotein complex is responsible for the staining reaction and for the absorption in the ultraviolet. But the latter method does not permit a distinction between DNA and RNA components.

There is, however, a reaction which does distinguish between these two types of nucleic acids. It has proved to be exceedingly useful in characterizing the nucleic-acid distribution in cells. This reaction is known as the Feulgen reaction. It depends upon the presence in nucleic acids of linkages which are differentially susceptible to hydrolysis. Using appropriate conditions, dilute HCl will hydrolyze certain sugar-base linkages in DNA but not in RNA. The hydrolyzed sugars of DNA are thus converted to reducing groups resembling aldehydes. These can be labeled by coupling them to a suitable colored reagent which reacts with aldehydes. Each single aldehyde group formed by hydrolysis is then tagged with a colored chromophore. This permits the DNA to be distinguished sharply from RNA. This useful reaction can be used quantitatively.

There is another important approach to the problem of distinguishing the localization in cells of RNA and DNA. This involves the use of specific enzymes. The enzyme, ribonuclease, will depolymerize RNA but not DNA. If RNA is removed from the nucleus by action of this enzyme, a loss of dye-binding capacity may appear in certain structures. The material originally present which was removed by the ribonuclease is thus shown to be RNA. It is also possible to remove DNA by deoxyribonuclease, and thus to ascertain which dye-binding component contains DNA. Still another method, often less reliable and less rigorously specific than the others, can be used to stain DNA and RNA in contrasting colors in the same cell. These three methods —the Fuelgen reaction, the use of specific enzymes, and the use of differential staining—permit one to distinguish cytochemically between DNA and RNA and provide information about the localization of these substances within the cell.

It turns out that DNA is found characteristically in the chromosomes and frequently as a jacket surrounding the nucleolar material. RNA is abundant in the central portions of the nucleoli and small amounts can be found within the chromosomes or scattered about in the nucleus.

If a large number of nuclei from a given species are examined by the Feulgen or some other suitable method and the amount of DNA in the individual nuclei is measured by absorption microspectrophotometry, as carried out by Alfert and by Swift, a small number of cells (for example, sperm cells) display a certain unit quantity of DNA. However, the majority of cells contains approximately twice this unit amount, while a third group of cells yields values clustering around four times the unit amount. In certain organisms—in the human liver, for example—one may find nuclei with up to eight times the unit value. From this, it derives that the amount of DNA in individual nuclei tends to occur in integral multiples of some unit quantity.

The chromosomes in dividing cells can often be counted with precision. In most higher animals, spermatozoa, the spermatids, and fully mature ova are found to contain a certain number of chromosomes. This number is called the "haploid number," and the chromosomes making up this number comprise a single set of chromosomes. These cells contain a single unit quantity of DNA, as mentioned earlier. The great majority of cells in most multicellular organisms contains twice the number of chromosomes characterizing a single set, and contains twice the unit quantity of DNA. Such cells are said to contain the diploid number of chromosomes. In diploid cells, most of the chromosomes occur in homologous pairs. Each member of a pair resembles its

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

mate closely, but morphological and genetic differences between the members of a homologous pair are often found. One may also find cells with four, six, eight, or some other small integral multiple of the haploid number of chromosomes, and with the same multiple of the unit amount of DNA found in haploid cells. Such cells are called tetraploid, hexaploid, or octaploid, respectively. In mammalian liver, one can find many tetraploid and some octaploid cells in a population which is predominantly diploid.

Each set of chromosomes contains a single set of genes. Thus, in diploid cells, many genes are represented twice, once in each member of a homologous pair. Such cells are said to be homozygous for the characteristics represented by genes which occur in identical form twice in each cell. But some gene loci are not identical in each member of a homologous pair of chromosomes. Such genes are represented only once in each cell, which is then said to be heterozygous with respect to the characteristic represented in those dissimilar loci. These data have contributed to the concept that there is a definite amount of DNA characteristic of each chromosome set, each chromosome, and each gene locus.

Some organisms live very well through most of their life cycle with haploid numbers of chromosomes in each nucleus. The fungus, *Neurospora*, for example, which has figured extensively in genetic studies, has a single set of chromosomes in each nucleus throughout long phases of the life cycle. The diploid phase may be very brief. In most metazoa, however, the diploid phase predominates.

One occasionally finds numbers of chromosomes that are *not* even multiples of the haploid number. Such cells are called "aneuploid" and contain amounts of DNA which are not integral multiples of the amount of DNA in haploid cells. Relevant examples are provided by certain cancers and leukemias. As a rule, the presence of aneuploidy and of morphologically abnormal chromosomes is interpreted as evidence for some genetic abnormality.

Levinthal (p. 227) remarks that information from genetics has led to the concept that the genetic carriers are arranged in linear sequence along some structure. The identification of the chromosome as this structure marked a major advance in the history of cytogenetics. Correlations between genetic and morphological data were most successfully worked out initially in *Drosophila*, where the chromosomes of the salivary glands are large and where genetic studies are readily carried out because of the short life cycle of the fly. Similar studies correlating genetic behavior with morphological abnormalities of chromosomes have now been carried out in a number of different forms. However, much of the corresponding work in virus and bacterial genetics is carried out conceptually, without direct visualization of chromosomes. The genetic data are used to construct a linear sequence, but no attempt is made ordinarily to correlate this sequence with morphological features of chromosomes—one important reason being that chromosomes have not been incontestably visualized in bacteria or in viruses. In these cases, the structural concept of linear sequence of units depends entirely upon genetic evidence.

At present, most of the important concepts of nuclear structure have been derived from genetic studies, from chemical analyses, and from examination with the light microscope. When combined ingeniously, these approaches have proved to be very powerful.

The electron microscope has yielded some additional information. It is evident that the intermitotic nucleus is surrounded by an envelope which consists of two unit membranes, an inner one and an outer one. These two membranes are joined to each other at certain sites. The lines of junctions surround and define pores several hundred Ångströms in diameter, perforating the nuclear membrane. Through these pores, the nuclear and cytoplasmic matrices communicate directly. The outer membrane of the nuclear envelope may also be continuous with membranes in the cytoplasm. The interior of the nucleus, however, is free from membranes, and is thus in sharp contrast to the cytoplasm, where membrane structures are frequent and often densely packed.

Electron micrographs usually show a rather irregular accumulation of granular material within the intermitotic nuclei. The granules presumably represent DNA and RNA, combined with protein. Very fine intranuclear helical threads have been described by some authors, but these are not seen with clarity and their significance is difficult to assess. No one has yet recognized structures in the nucleus corresponding to the nucleic-acid helices as studied by Hall and Doty, with the electron microscope, and pictured elsewhere in this volume (p. 107). One can find nucleoli in electron micrographs. They appear as rather irregular accumulations of granules more densely crowded together than elsewhere in the nucleus.

Mitotic cells show no nuclear envelope. The chromosomes are recognizable as dense accumulations of granular material without any membranous investment.

Thus, it appears that direct electron microscopy of the nucleus of ordinary cells has provided little information which is satisfying in the light of the physiological importance of the nucleus and its contents.

In certain specialized types of cells, organized structures have been detected with the electron microscope which are not characteristic of cells in general. Thus, in some sperm cells, dense striations running the length of the nucleus have been seen. Moses has detected more-delicate organized structures in certain spermatocytes. These appear as long thread-like elements with delicate side chains extending laterally. Efforts have been made to associate the filaments with some phase of chromosome structure. But, at the present time, it

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

would be hazardous to do more than point out that the dimensions of the threads of the side loops are not inconsistent with those of a nucleic-acid helical chain.

Although the electron microscope shows that the nucleus contains particles which closely resemble the RNP particles of the cytoplasm, it reveals relatively little concerning the mechanism of transfer of nucleic acid from nucleus to cytoplasm. Such a transfer would provide a means whereby the nucleus could transmit information to the cytoplasm. Yet one can see images which suggest transition stages in the course of outward movement of RNA from the nucleus. One may occasionally see accumulations of particles in the cytoplasm which very closely resemble the nucleoli. Similar appearances observed with the light microscope have led to the view that in some cells whole nucleoli may be discharged from the nucleus carrying large packets of RNA to the cytoplasm. On the other hand, it may be that much of the RNA escapes into the cytoplasm through the pores in the nuclear envelope. Another mechanism whereby RNA could pass from nucleus to cytoplasm involves a binding of the newly found RNA to the inner nuclear membrane. The latter then would flow into the cytoplasm, carrying the bound particles with it. Such a mechanism, however, is not well documented.

We are forced, then, to conclude that the task of correlating the fine structure of the nucleus with its function and with the chemical and genetic evidence at our disposal is largely before us.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

REVIEWS OF MODERN PHYSICS VOLUME 31, NUMBER 2 APRIL, 1959

37

Fine Structure of Cytoplasm: The Organization of Membranous Layers

FRITIOF S. SJÖSTRAND

Karolinska Institutet, Stockholm 60, Sweden

WHEN climbing the ladder of levels of organization of living matter, one finds nowadays that at least indications of steps are present along the whole ladder. Some steps may be weak and might break when exposed to too critical attention. Light microscopy has furnished information regarding a great number of subcellular structural elements, the so-called cell organelles, as well as of cytoplasmic differentiations of a more diffuse character. The most impressive of the cell organelles which are localized in the cytoplasm are the mitochondria and the Golgi apparatus. Basophilic regions of the cytoplasm represent more-diffusely distributed parts which long ago were called the ergastoplasm by Garnier. The reason for this term was the idea that these regions were especially active parts of the cytoplasm. In the exocrine pancreas cell, a filamentous or lamellar structure was observed within the basophilic regions of the cytoplasm, "die Basalfilamenten" of Heidenhain, or "die Basallamellen" of Zimmermann. These various components of the cytoplasm were considered immersed in the ground substance of the cytoplasm, a component which was assumed to lack any higher degree of specialized organization.

The submicroscopic organization of the regularly occurring cell organelles was completely unknown, however. A few cell components, such as the myelin sheath of peripheral nerve fibers and the outer segments of the retinal receptors of the vertebrate eye, were assumed to consist of alternating layers of lipid and protein molecules. This assumption was based upon polarization-optical data of these strongly birefringent components.

Electron microscopy has made possible the fairly close analysis of the submicroscopical organization of these various components. In some cases, the combination of electron-microscope data with polarization-optical and x-ray diffraction data has made it possible to propose models containing some features of the molecular architecture of these components. In this latter respect, the results can be considered rather crude.

On the other hand, electron microscopy has not revealed any definite new structural component of the cytoplasm. It has helped in bridging the gap between the molecular and the organizational levels of the cell organelles by making supramolecular elements available for direct observation.

The most striking feature of the ultrastructural organization of the cytoplasm is the frequent occurrence of membranous components in various cell organelles, as well as in the ground substance of the cytoplasm. Membranes of various dimensions and different organization appear to represent a common and basic principle of organization in the cytoplasm.

Some examples of such membranous elements are presented here, together with some arguments which justify an interpretation of some of the structural patterns as reflecting a certain molecular structure of the membranes. Also, the probability that the observed structural patterns represent preformed patterns existing in the intact living cell is discussed. Finally, some ideas are presented regarding the functional significance of the various membranous elements.

When evaluating the structural patterns observed by means of the electron microscope, it is wise to account for a certain degree of deformation introduced by preparatory techniques. When comparing two patterns, sometimes one pattern easily might be imagined as derived from the other as a result of disorganization or deformation. In such a case, the second pattern is selected as probably more directly related to the *in vivo* pattern. When a certain variation in the patterns is observed, the most frequently occurring pattern is selected as the most representative.

The chapter on the biochemistry of mitochondria by Lehninger (p. 136) makes it quite natural to start this survey with some electron micrographs of mitochondria. For this purpose, mitochondria may be chosen from almost any type of cell. Let us choose those of the retinal receptors of the eye. The first figure shows a schematic drawing of the rod type of receptor in the guinea pig retina. These receptors are segmented cells with each segment structurally organized in a characteristic and different way (Fig. 1). The outer segments which are most remote from the pupil of the eye contain the photochemically active molecules, such as rhodopsin in the rod cells. Here, the primary reactions in converting electromagnetic energy into chemical energy take place. The inner segment contains all of the mitochondria of the receptor and can be looked upon, therefore, as the energy-generating center of the cell.

The rod and cone fibers connecting the inner segment with the synaptic bodies are organized like unmyelinated nerve fibers. The vitreous end of the receptor cell forms the synaptic body where the synaptic contacts between receptors and nerve cells are located.

Figures 2 and 3 picture the inner segment, *IS*, of a cone cell in the perch retina. Most of the inner segment consists of a dense aggregation of closely packed mitochondria which together form the so-called ellipsoid.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

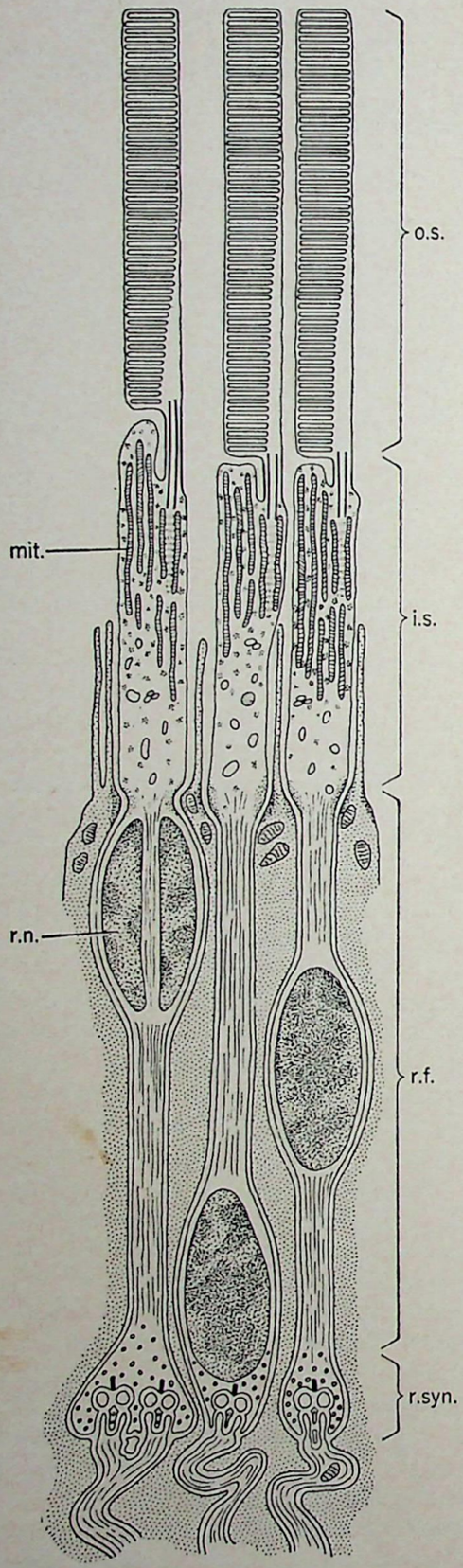


FIG. 1. Schematic drawing of retinal rods in the guinea pig eye. *o.s.*, outer segment; *i.s.*, inner segment; *r.f.*, rod fiber; *r.syn.*, rod synaptic body; *mit.*, mitochondria; *r.n.*, rod nucleus [from F. S. Sjöstrand, Intern. Rev. Cytol. 5, 455 (1956)].

Each mitochondrion is bounded by a surface membrane, and a great number of inner membranes or platelets are oriented in a fairly parallel fashion in the interior. All of these membranes appear to be triple layered.[1–4] It is striking to note the rather constant spacing of the layers of these mitochondria membranes.

In 1952, when we first presented our structural model of mitochondria, Palade[5] proposed a different model consisting of a single-layered surface membrane which showed local infoldings which he called "cristae mitochondriales." These cristae left a central space in the mitochondrion free. After confirming our observations on the triple-layered character of the surface membrane, Palade changed his model by enveloping his original model in a peripheral opaque layer. He assumes[6] that this layer represents the surface membrane of the mitochondrion and that an outer space or chamber extends between the two opaque layers located at the surface of the mitochondrion. This space is continuous with spaces which extend in the cristae. An inner mitochondrion chamber forms a continuous central space. This interpretation differs from our own.

Our interpretation of the mitochondrial pattern is based upon certain observations that we had made on some typical lipoprotein systems—namely, the outer segments of retinal receptors[7,8] and the myelin sheath of peripheral nerves.[9,10] In both cases, polarization-optical data, and, in the myelin sheath, x-ray diffraction data had permitted the proposal of models showing alternating layers of lipid and protein molecules. X-ray diffraction data revealed some dimensions for the thickness of the layers in the myelin sheath and of the double layer of dried, mixed nerve lipids. For the latter, a value of 67.4 A was obtained.[11]

A discussion of the observations made on the outer segments of retinal receptors gives a necessary background for our interpretation of the mitochondrial pattern. In electron micrographs of sections (Figs. 4–6) oriented parallel to the long axis of the outer segments of rods and cones, alternating light layers and pairs of opaque layers are seen, the two opaque layers of a pair being fused along their rims. They thus bound a light interspace which, in guinea pig rods, has a thickness of 70 to 80 A. These pairs of layers form triple-layered disks, the total thickness of which varies from 100 to 250 A with the species and with the type of receptor, but remains constant for any one type of receptor. These disks are the unit structure of the outer segments.

The outer segments are 40% lipid in content.[12] More than 80% of these are phospholipids. The polarization-optical data have revealed that presumably the lipid molecules are oriented parallel to the long axis of the outer segment, and that the protein molecules form transversely oriented layers.[13]

When fragmenting isolated outer segments, it is possible to obtain round disks of varying thicknesses and also fragments of such disks.[7] The thicknesses of the disks isolated from the guinea pig retina are multiples of 140 A—that is, multiples of the thickness of the triple-layered unit disk which can be observed in sections and which measures 140 A in thickness in the

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

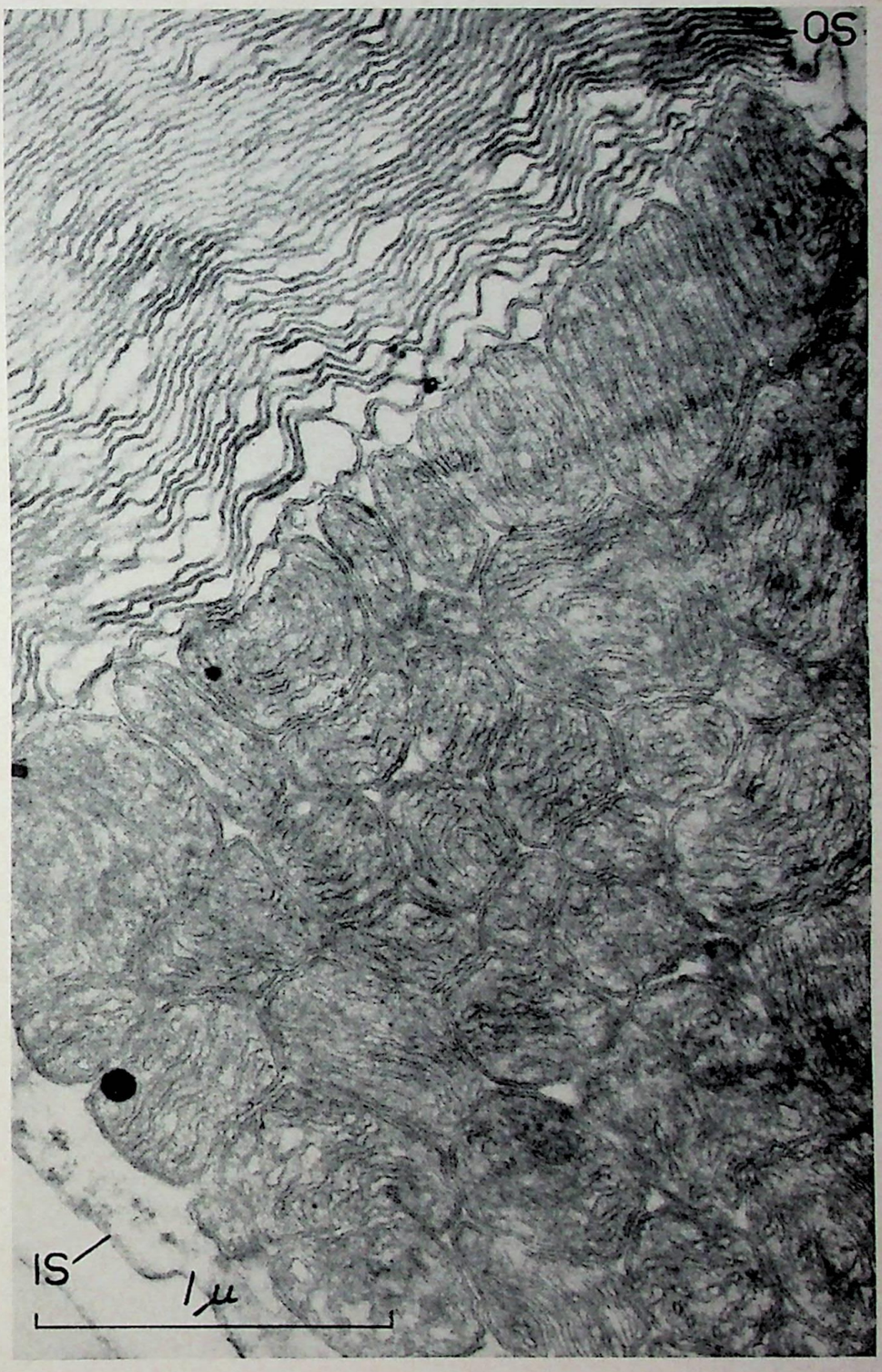


FIG. 2. Survey picture of the boundary between the outer, *OS*, and inner, *IS*, segments of a cone cell in the perch retina. Most of the inner segment consists of a dense aggregation of mitochondria [from F. S. Sjöstrand, *Ergebnisse der Biologie* (Springer-Verlag, Berlin, 1958), Vol. XXI, p. 128]. ×44 000.

guinea pig retina. The thickness of the thinnest membrane fragments is about 30 A. These fragments are osmiophilic. The thickness of the osmiophilic layers of the unit disks as observed in sections through guinea pig retinas is 30 to 40 A. There seem to be no doubt that these components in fragmented and in sectioned material are identical. The 140-A thick disks which were obtained by fragmentation could be demonstrated to consist clearly of two thin membranes, and under favorable conditions they could be split completely into two thinner components with a thickness of 70 A at the edge.

The fact that the thicknesses of the various disks obtained by fragmentation represent multiples of 140 A makes it justifiable to conclude that, when the outer segments are fragmented in piles of unit disks of different thicknesses and when these piles are dried on a supporting film, the interspaces between the triple-layered disks contribute insignificantly to the thicknesses of the piles. This means that these interspaces

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

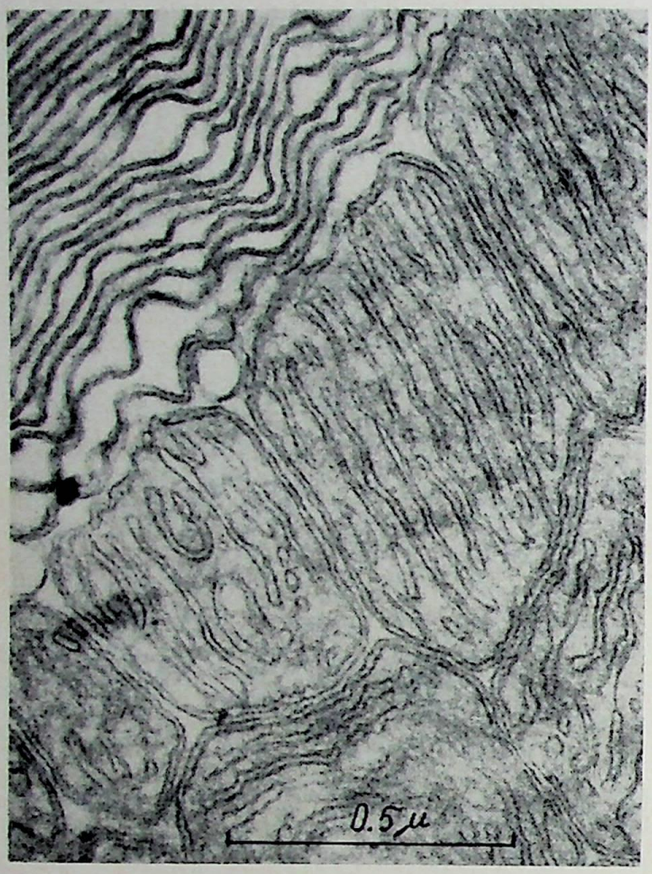


FIG. 3. Higher magnification of a region at the border between the inner and outer segments of a cone cell in the perch retina. Several mitochondria are observed densely aggregated. The triple-layered surface membrane and inner membranes show certain similarities to the triple-layered disks of the outer segment [from F. S. Sjöstrand, *Ergebnisse der Biologie* (Springer-Verlag, Berlin, 1958), Vol. XXI, p. 128]. ×65 000.

are filled mainly with an aqueous, ionic medium containing little or no lipids. The lipid molecules, therefore, are located presumably in the triple-layered disks. Their localization (Fig. 7) in the less-opaque space bounded by the 30- to 40-A thick osmiophilic layers was assumed because the dimension of this space, 70 to 80 A, could well accommodate a double layer of lipid molecules. Furthermore, the osmiophilic layers had shown a rather remarkable tensile strength in the fragmentation experiments which appeared to point to a protein nature. The total volumes of the opaque layers and of the spaces surrounded by these layers are estimated to be about equal. This observation is in accordance with the determined figure of 40% for the concentration of lipids in percent of dry weight.

The similarity of the mitochondrial pattern and that of the outer segments of retinal receptors is rather striking. Values as high as 40% were reported for the lipid content of mitochondria, and they showed positive birefringence after freeze-drying,[14] which could be assumed to be due to oriented lipid molecules. The triple-layered membranes of the mitochondria were interpreted as representing two protein layers sandwiching a double layer of lipid molecules (Fig. 8). This interpretation is based upon a series of evaluations of indirect evidence and upon a generalization which well might be erroneous.

The myelin sheaths of peripheral nerves also show a characteristic layered structure. Furthermore, peripheral nerves can be fragmented into thin membranous fragments.[15,16] In this instance, polarization-optical data, as interpreted by Schmidt,[13] revealed that the lipid molecules were oriented radially and that protein molecules formed concentric layers around the axis cylinder. X-ray diffraction data by Bear, Palmer, and Schmitt[11] were interpreted as showing a repeat period of 171 A in amphibian nerves, and 185 A in mammalian nerves. The repeat period was interpreted as containing two bimolecular leaflets of lipids, each about 67 A thick, and a layer of proteins about 25 A thick. A concentrically layered structure could be observed in the electron micrographs, obtained in 1952,[9,10] of osmium-fixed, sectioned peripheral nerves. Opaque layers about 25 A thick are arranged concentrically in the myelin sheath with a definite periodicity. In each period, a fainter opaque line is seen to halve the main period. The mean periodicity was estimated at 120 A in the electron micrographs of 20 nerve fibers.

Although the x-ray diffraction and the electron-microscope data did not coincide as to the length of the period, the over-all scheme predicted from the former data appears to be pretty well confirmed by the latter. From the dimensions, it was concluded that the opaque layers correspond to the protein layers, and the less opaque layers to bimolecular leaflets of lipids. The recent extensive work by Fernández-Morán and Finean[17] on this problem seems to support this interpretation.

A membranous component which is frequently present in the basophilic regions of the cytoplasm is discussed next. These membranes are particularly abundant in the excretory pancreas cells, the salivary gland cells, the plasma cells, and other types of cells in which a somewhat intense protein synthesis takes place.

In the exocrine pancreas cells (Fig. 9), most of the ground substance of the cytoplasm is filled with this type of membrane.[1,18] In osmium-fixed material, the membrane is easily identified by the numerous opaque particles that are attached to one surface. These particles have an average diameter of 150 A. Their form varies; the particles are more or less irregularly, angularly shaped. The membrane to which these particles are attached consists of an osmiophilic component which is about 40 A thick. These membranes are arranged in pairs and a pair bounds a narrow space or, as in the thyroid epithelium, rather large, irregularly shaped spaces. They thus appear to divide the cytoplasm into two different parts or compartments.[19]

Opaque particles of identical appearance, as those attached to the membranes, frequently occur free in the cytoplasm. Palade and Siekevitz[20] clearly demonstrated that the opaque particles are responsible for the RNA content of the microsome fraction from pancreas tissue. The terms microsomes, ribosomes, or microsomal

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

FIG. 4. Longitudinal section through the outer segment of a retinal rod cell from the perch eye showing the uniformly layered structure of the outer segment. ×59 000.

particles now seem to be used to classify this component which represents only a minor part of the classical microsome fraction.

At this point, it seems justifiable to make some comments on what is known about the fixation and embedding artifacts that are introduced when fixing

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

FIG. 5. Longitudinal section through a retinal-rod cell of the perch eye showing the triple-layered disks, which represent the elementary component of the outer segment [from F. S. Sjöstrand, *Ergebnisse der Biologie* (Springer-Verlag, Berlin, 1958), Vol. XXI, p. 128]. ×82 000.

and embedding labile living matter. The tissue is fixed in a solution of osmium tetroxide, dehydrated in a graded series of concentrations of ethyl alcohol, and then transferred to methacrylate which is polymerized before the tissue can be sectioned.

At the beginning, we were very much concerned about this problem and, therefore, we intentionally chose to analyze such structural components about which a little was known from an ultrastructural point of view through polarization-optical and x-ray diffraction analysis. In the outer segments of retinal receptors and in the myelin sheath, patterns in complete harmony with polarization-optical data obtained from fresh, unfixed material were observed by means of electron microscopy, as mentioned in the foregoing. Regarding the exocrine pancreas cells, there were no data available regarding birefringence of the cytoplasm *in vivo*. Living exocrine pancreas cells in mice were analyzed in polarized light and we found that the cytoplasm was birefringent.[1] When analyzed, this birefringence proved to be negative with the axis oriented perpendicularly to the cell membrane, a result that could fit in very well with the electron-microscope picture.

When the direct study of living material is excluded, the most rational way of checking the reliability of such a preparatory technique is to change the technique so radically that it appears unlikely that the two techniques could produce identical distortions. The second preferred preparatory technique is then preservation by means of freeze-drying. In fact, we proceeded in the opposite direction and studied frozen-dried material first[1] and later checked the observations made on exocrine pancreas cells by means of osmium fixation and birefringence analysis.

The freeze-drying technique is a rather delicate procedure with many pitfalls. The observations made on frozen-dried, sectioned material, stained with phosphotungstic acid, were rather surprising.[21] Similar structural patterns could be observed in the ground substance of the cytoplasm as well as in the mitochondria, with the remarkable exception that the patterns appeared as negative patterns, as compared to those obtained after osmium fixation (Figs. 10 and 11). These results mean that these patterns probably are preformed and exist in the living cell. At the same time, however, they may make it necessary to re-evaluate the interpretation of the molecular structure assumed to be responsible for these patterns.

One feature of the frozen-dried specimen must be noted. The so-called RNA particles do not show up.

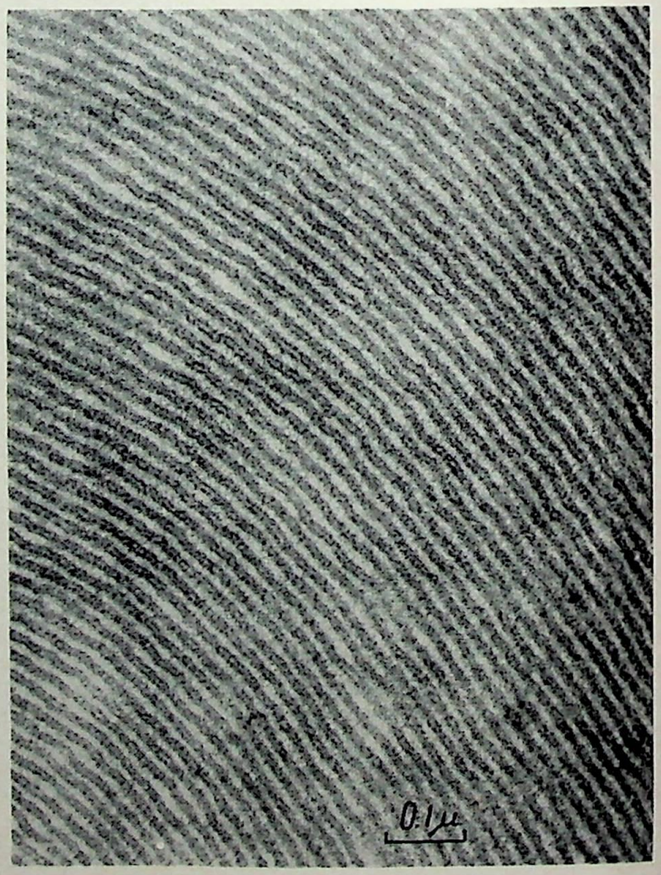


FIG. 6. Section oriented perpendicularly to the triple-layered disks of the outer segment of a cone cell in the perch eye. Notice the tendency to a uniform blackening of the whole triple-layered disk due to a staining of the middle layer [from F. S. Sjöstrand, *Ergebnisse der Biologie* (Springer-Verlag, Berlin, 1958), Vol. XXI, p. 128]. ×87 000.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

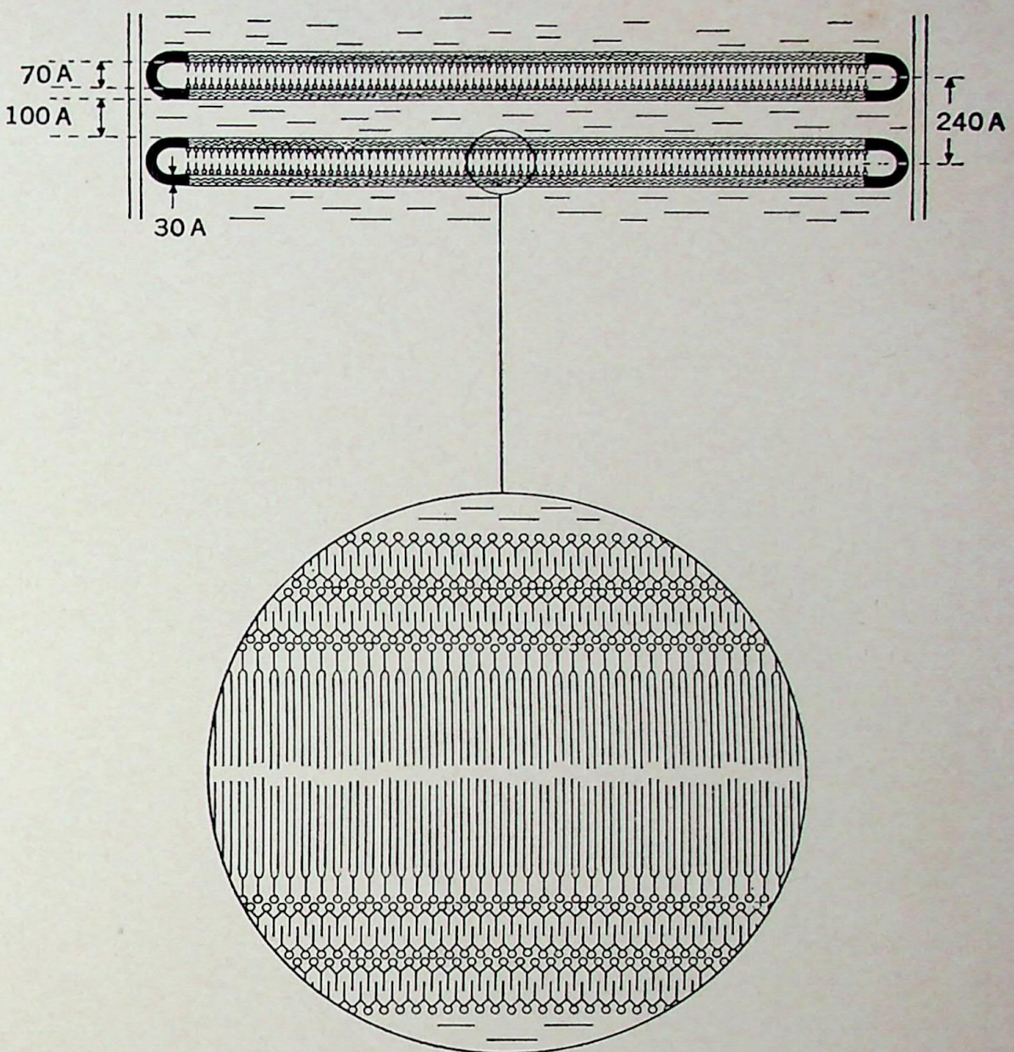


FIG. 7. Schematic drawing to illustrate a proposed interpretation of the localization and orientation of lipid and protein molecules in the triple-layered elementary disks of the outer segments of retinal receptors [from F. S. Sjöstrand, *Ergebnisse der Biologie* (Springer-Verlag, Berlin, 1958), Vol. XXI, p. 128].

The cytoplasm between the cytoplasmic membranes appears homogeneous even after long periods of staining. One cannot exclude, therefore, the possibility that these particles are artifacts formed through an aggregation or precipitation in connection with the preparatory procedure. This possibility has to be tested more extensively than has been done to date.

Even if some doubts can be raised regarding the particulate form of the RNA-protein components in the cytoplasm of intact living cells, there is little or no doubt that these particles indicate the approximate location of RNA in the cytoplasm. The firm connection of the particles with the cytoplasmic membranes, on the other hand, can be questioned as due to secondary absorption. This would explain why such particles are attached to the outer opaque layer of the nuclear membrane in cells where the RNA is located close to the nucleus. Furthermore, it explains why the particles are absent when the cytoplasmic membrane is located very closely to mitochondria where the amount of RNA easily can be assumed to be inadequate for the formation of such RNA particles.

Another structural component of the cytoplasm is the so-called *Golgi apparatus*. This component has been interpreted by light microscopists as an apparatus involved in secretory mechanisms, and it has been extensively studied in glandular cells. The secretory products were assumed to be either produced or accumulated in this region.

Electron microscopy has revealed another type of cytoplasmic membrane as the main component of the Golgi apparatus.[22–24] These membranes are arranged in pairs also and bound a space that can be either very narrow, about 100 A or less in width, or very wide appearing as large vacuolar, irregularly shaped spaces (Figs. 12 and 13). The membranes consist of one osmiophilic component, 60 A thick. Presumably, a less osmiophilic component separates the membrane pairs, because the space between the pairs frequently is very constant, measuring about 60 A in width.

It is rather striking that the zymogen granules which contain the secretory products of the pancreatic cells show a definite topographic relationship to the Golgi apparatus. The zymogen granules can be divided into two types—the fully-developed granules and the precursor granules. A complete series of intermediate-type granules bridges over from the one type to the other. The fully developed zymogen granules are spherical

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

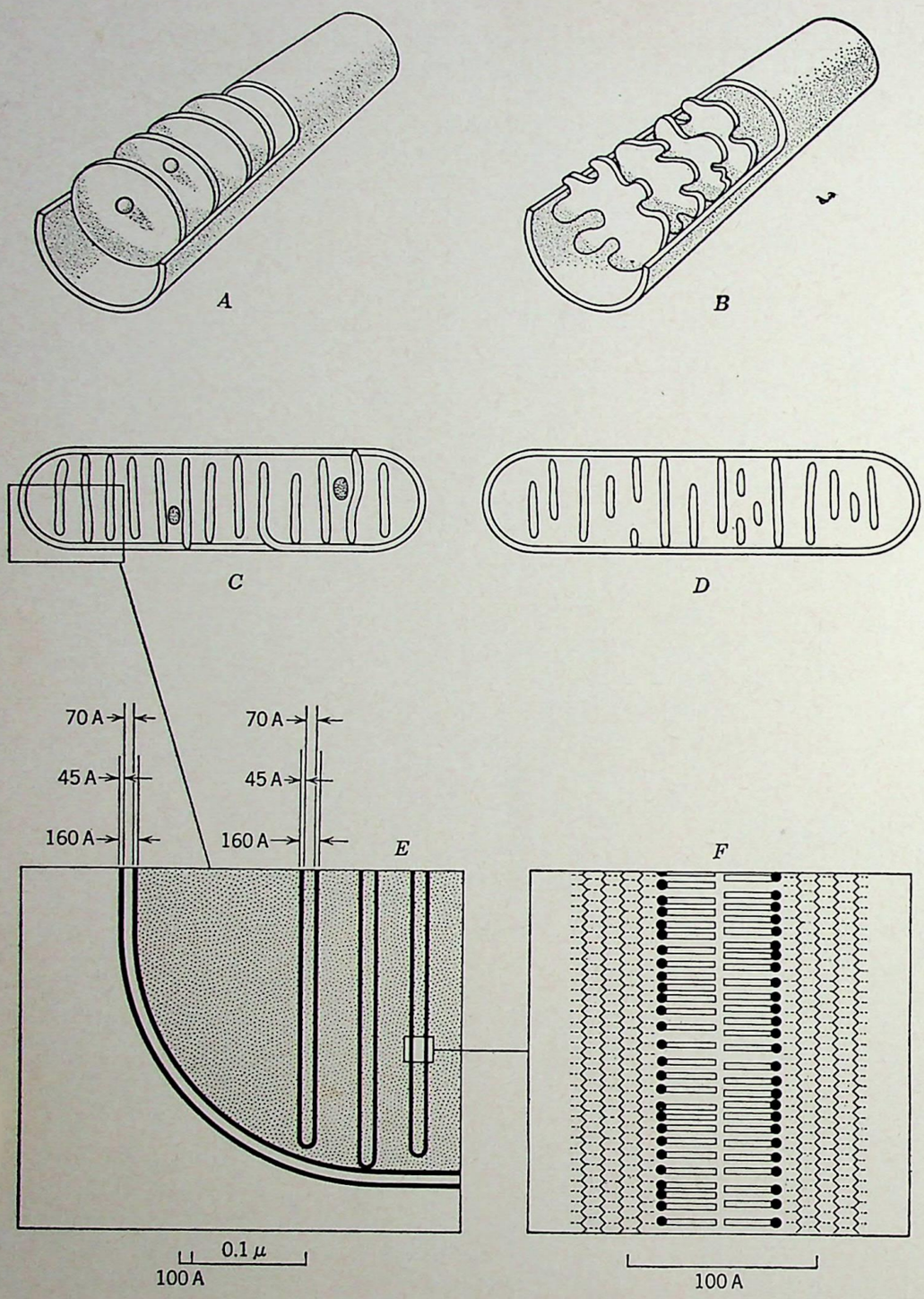


FIG. 8. Schematic drawing which illustrates an interpretation of the three-dimensional appearance of the mitochondria and of the molecular architecture of the mitochondria membranes. The three-dimensional representation *A* and *B* is based on an interpretation of the patterns observed in sections which can be, for instance, like the ones shown in *C* and *D*. *E* gives some commonly observed dimensions of the mitochondrial membranes, and *F* a scheme of the proposed arrangement of lipid and protein molecules in the membranes [from F. S. Sjöstrand, in *Fine Structure of Cells*, Proc. Symp. VIII Cong. Cell. Biol., Leiden, 1954 (Noordhoff, Ltd., Leiden, 1955), p. 16].

and of high density, and are bounded by a surface membrane about 50 A thick. The precursor granules are irregularly shaped and of lower opacity, but are bounded by a surface membrane which shows out-pocketings and bulgings (Fig. 13). The precursor granules are found only in connection with the Golgi apparatus where the fully developed zymogen granules are either rare or completely absent. Furthermore, there is a definite polarity of the Golgi apparatus with the precursor granules arranged on the apical side of the Golgi membranes. On the basal side, the type of cytoplasmic membranes described earlier reaches close to the Golgi membranes. A zone of granulated cytoplasm, however, separates the territories of these two types of membranes.

Turning to the problem of the functional significance of these various kinds of membranes, the mitochondrial membrane is considered first. The article by Lehninger (p. 136) introduces the present concept regarding the importance of the structural organization of the mitochondria. The biochemical data point to a structural factor as important in connection with coupled phosphorylation. The electron transport seems to depend upon an organization of macromolecular particles. The

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

FIG. 9. Cytoplasmic membranes in the basophilic cytoplasm of exocrine pancreas cells. Notice the opaque 150-Å particles attached to the one surface of the membranes which are arranged in pairs. In the lower part of the picture a cell boundary represented by two mutually parallel lines, and below the boundary a mitochondrion [from F. S. Sjöstrand and V. Hanzon, Exptl. Cell Research 7, 393 (1954)]. ×108 000.

smooth coordination of various enzymatic reactions was imagined as being the result of a topographic arrangement of the enzyme molecules according to certain patterns. The membranes of the mitochondria were assumed to be composed to a great extent, or altogether, of mitochondrial enzyme molecules oriented and spread out in layers in combination with layers of oriented lipid molecules.

The experiments done so far to isolate the membranous components of the mitochondria from the ground substance have been rather crude from a morphological point of view, but are interpreted as supporting this idea. The mitochondrial membranes are considered to be enzymatically active where the organization of the enzyme molecules in a plane increases the chances for a well-coordinated sequence of enzyme-substrate interactions. This organization is thought to be important for a rapid step-wise degradation of substrate molecules, and possibly for their step-wise synthesis.

One may assume that the synthesis of the secretory products in the exocrine pancreas cells takes place

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

FIG. 10

FIGS. 10–11. Sections through exocrine pancreas cells preserved by means of freeze-drying. The methacrylate section was stained with phosphotungstic acid before the examination in the electron microscope. The cytoplasmic membranes are visible as well as the mitochondria but the pattern is a negative one as compared to that shown in Fig. 2. The spaces between the membranes (Fig. 11) are uniformly filled with material in contrast to the lack of material, except the opaque 150-A particles, after osmium fixation. No such particles can be observed in these pictures. *Mi*, mitochondria; *CM*, cytoplasmic membranes; *V*, ice crystal vacuole [from F. S. Sjöstrand and R. F. Baker, J. Ultrastructure Research 1, 239 (1958)]. ×30 000 (Fig. 10); ×56 000 (Fig. 11) (Fig. 11 is on opposite page).

in steps. First, an enzyme-containing membrane is formed which is able to synthesize the secretory products (Fig. 14). This membrane is manufactured in the Golgi apparatus from material delivered by the cytoplasm as represented by the RNA particles. These particles swell, fuse, and mix with lipids, and extensive hydration results in the formation of vacuoles bounded by a thin membrane consisting of lipoproteins. The vacuoles collapse and pairs of membranes are formed which disintegrate into fragments forming the surface membrane of the precursor granules. The surface membrane then synthesizes the secretory products which are successively accumulated in the granule. Raw material can be delivered all of the time by the RNA-rich cytoplasm.

The foregoing is a wild hypothesis that could explain the morphological observations made so far. It is derived, however, from still pictures; the various stages

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

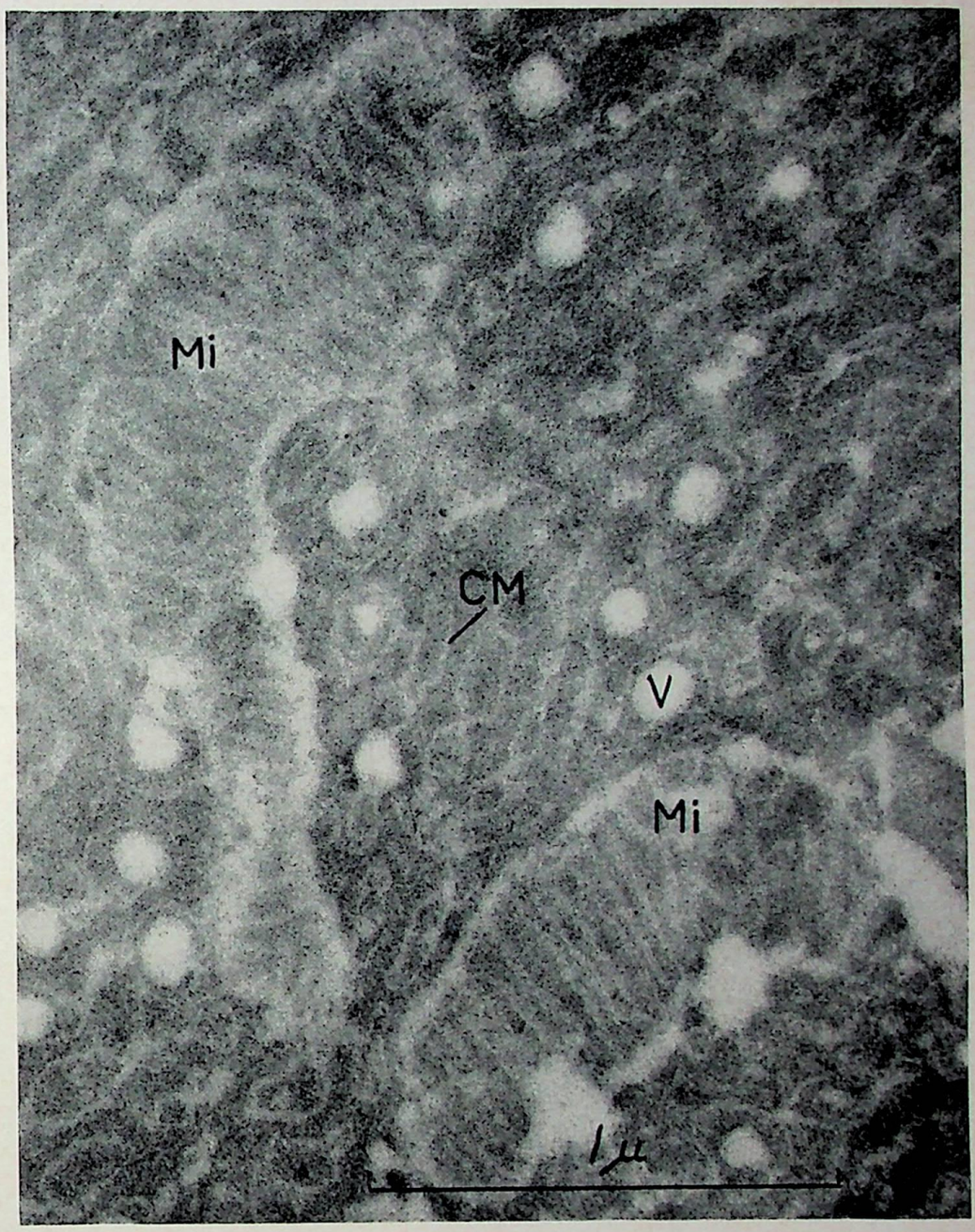


FIG. 11 (for legend, see Fig. 10).

have been selected subjectively from these pictures. Experimental evidence remains to be presented. One thing seems to be quite clear—namely, that the Golgi membranes can form vesicles or the membranes of secretory granules. This latter process is demonstrated better by the goblet cells in the intestinal epithelium. Holmberg,[25] in our laboratory, found that, after administration of diamox, an inhibitor for carbonic anhydrase, the Golgi apparatus in the ciliary epithelium of the eye partially disintegrated into vesicles and the whole cytoplasm became loaded with vesicles. These reversible changes could be described in a quantitative way and be correlated with a temporary 60% inhibition of the secretion of aqueous humor.

In certain types of cells, the plasma membrane appears extremely folded. It is not quite clear whether these folds are tight infoldings of the plasma membrane or whether they represent interdigitating ridges of the cell surfaces of adjacent cells. Zetterqvist,[26] in our laboratory, demonstrated that the plasma membrane can appear rather different at various parts of the surface of the same cell. In the intestinal epithelium (Fig. 15) the free surface of the cells forms a number of cylindrical processes, the so-called brush border. The plasma membrane bounding these processes appears as a 105-A thick triple-layered component (Fig. 16). Where two epithelial cells are in close contact, the plasma membrane appears as an osmiophilic layer 60 to 70 A thick. Between the two osmiophilic layers of the two adjacent epithelium cells, there is a light space with a very uniform thickness of about 100 A.

On the other hand, the plasma membrane which

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

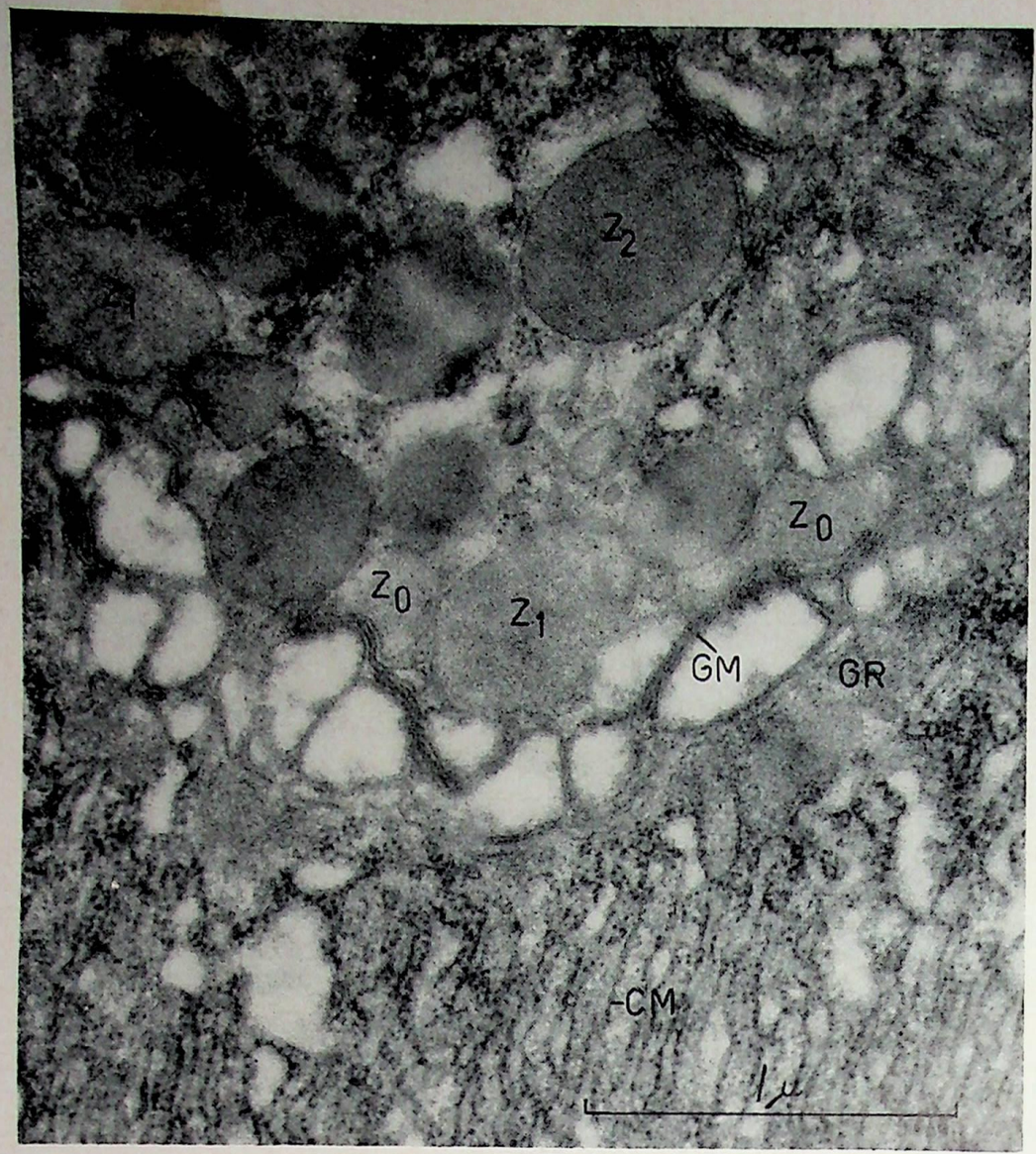


FIG. 12. A Golgi apparatus in an exocrine cell of the mouse pancreas. *GM*, Golgi membranes; *GR*, granules on the basal side of the Golgi apparatus; *CM*, cytoplasmic membranes; Z_0, prozymogen granules that appear to develop from the Golgi membranes; Z_1, prozymogen granules which can be interpreted as intermediate stages between Z_0 and Z_2, which are fully developed zymogen granules [from F. S. Sjöstrand and V. Hanzon, Exptl. Cell Research **7**, 415 (1954)]. ×44 000.

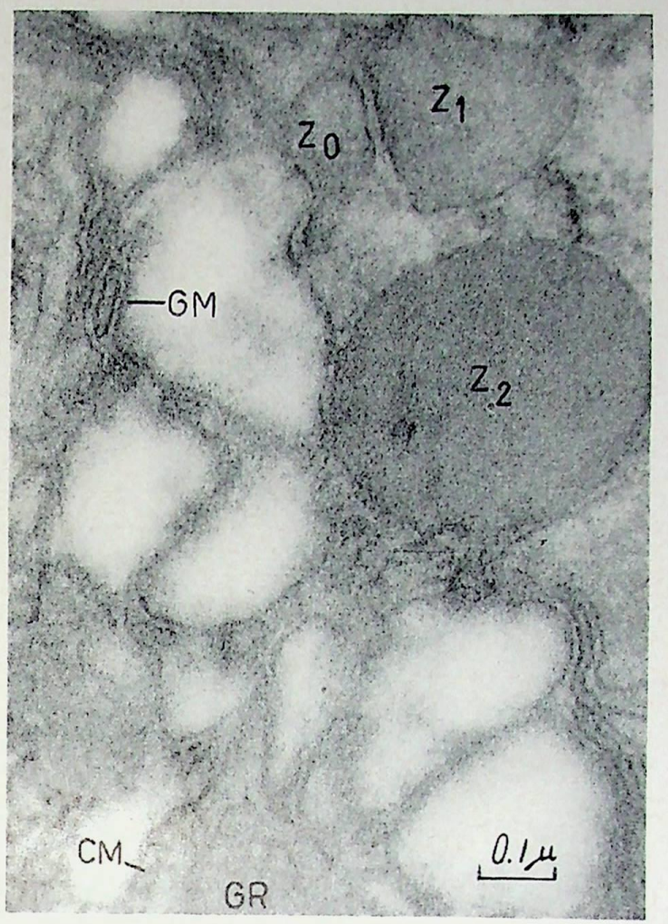


FIG. 13. Higher magnification of a part of Fig. 12. For explanation of indications, see Fig. 12. ×83 000.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

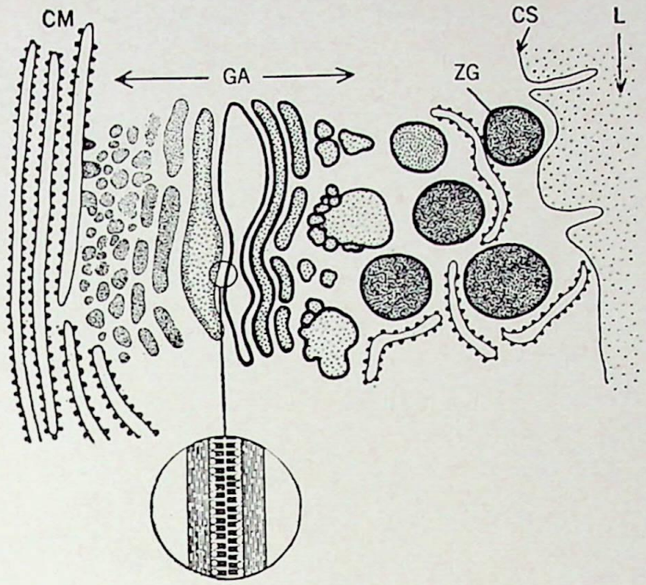


FIG. 14. Schematic drawing illustrating a hypothesis which aims at interpreting the functional significance of the Golgi apparatus of the excretory cells of the pancreas as an assembly line for membrane material with the power to synthesize the secretory products of the exocrine cells from raw material in the cytoplasm. *CM*, cytoplasmic membranes; *GA*, Golgi apparatus with formation of membranes from material delivered by the cytoplasmic membranes; *ZG*, zymogen granules formed from precursor granules in the Golgi apparatus; *CS*, cell surface facing the lumen of the excretory duct, *L* [from F. S. Sjöstrand, UNESCO Symp. Patterns of Cellular and Sub-Cellular Organisation, Edinburgh, 1957 (to be published.)]

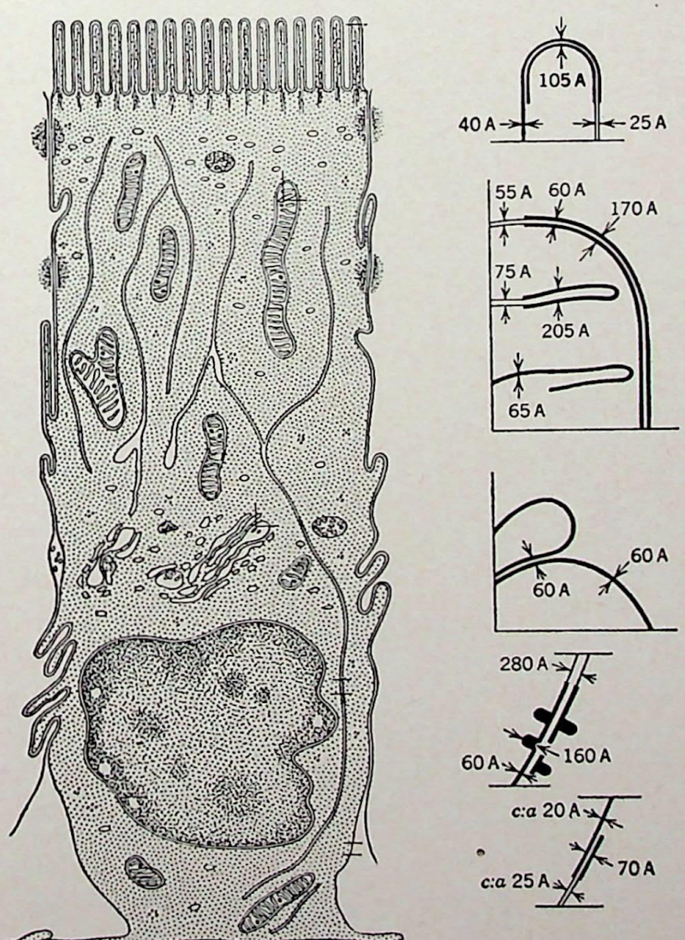


FIG. 15. Schematic drawing of a cylindrical epithelial cell lining the small intestine in the mouse. Notice the so-called brush border at the upper surface of the cell which faces the lumen of the gut [from H. Zetterqvist, *The Ultrastructural Organization of the Columnar Absorbing Cells of the Mouse Jejunum*. Thesis, Stockholm (1956)].

covers the surface of the cell facing the intercellular spaces and the basement membrane appears as a triple-layered structure which measures only 70 A in thickness. This structure consists of two opaque layers, each about 25 A thick, separated by a 20-A thick light interspace. A similar appearance has been reported recently by Robertson[27] for the plasma membranes of the Schwann cells. Robertson has interpreted this approximately 70-A thick triple-layered structure as representative of the whole plasma membrane (unit membrane) and which consists of two protein monolayers sandwiching a double layer of lipid molecules.

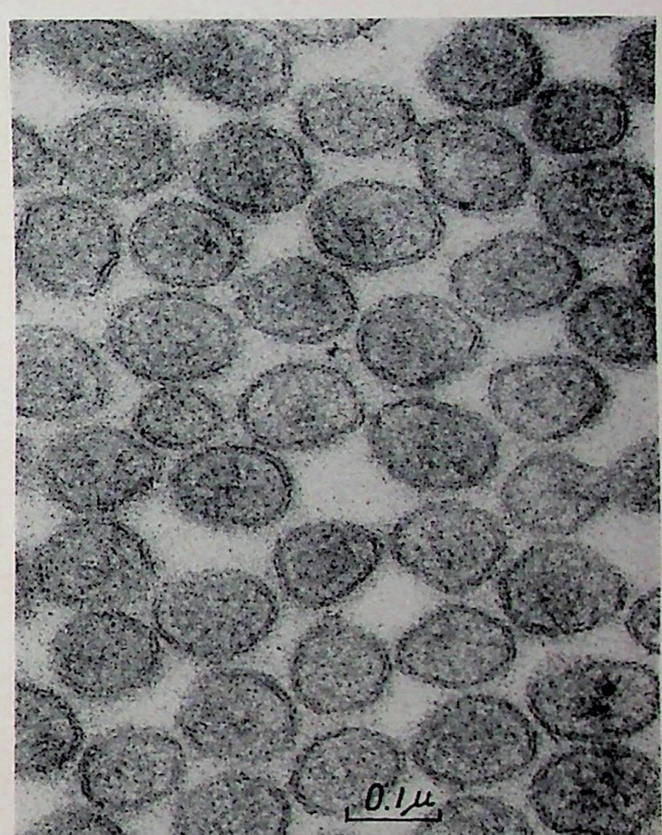


FIG. 16. Transversal section through the cylindrical processes of the brush border zone in an intestinal epithelial cell. The processes are bounded by a triple-layered surface membrane with an average total thickness of 105 A [from H. Zetterqvist, *The Ultrastructural Organization of the Columnar Absorbing Cells of the Mouse Jejunum*. Thesis, Stockholm (1956)]. ×103 000.

This interpretation is based on a great deal of evidence and it is discussed further by Schmitt (p. 455). It is assumed that the consistency of the width, about 100 A, of the light interspaces between two cells in close contact indicates the presence of some binding material, such as lipids or mucopolysaccharides.

Consider now the formation of membranous material in the outer segments of retinal receptors. It is seen, in the double cones of the perch retina, that the disks of

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

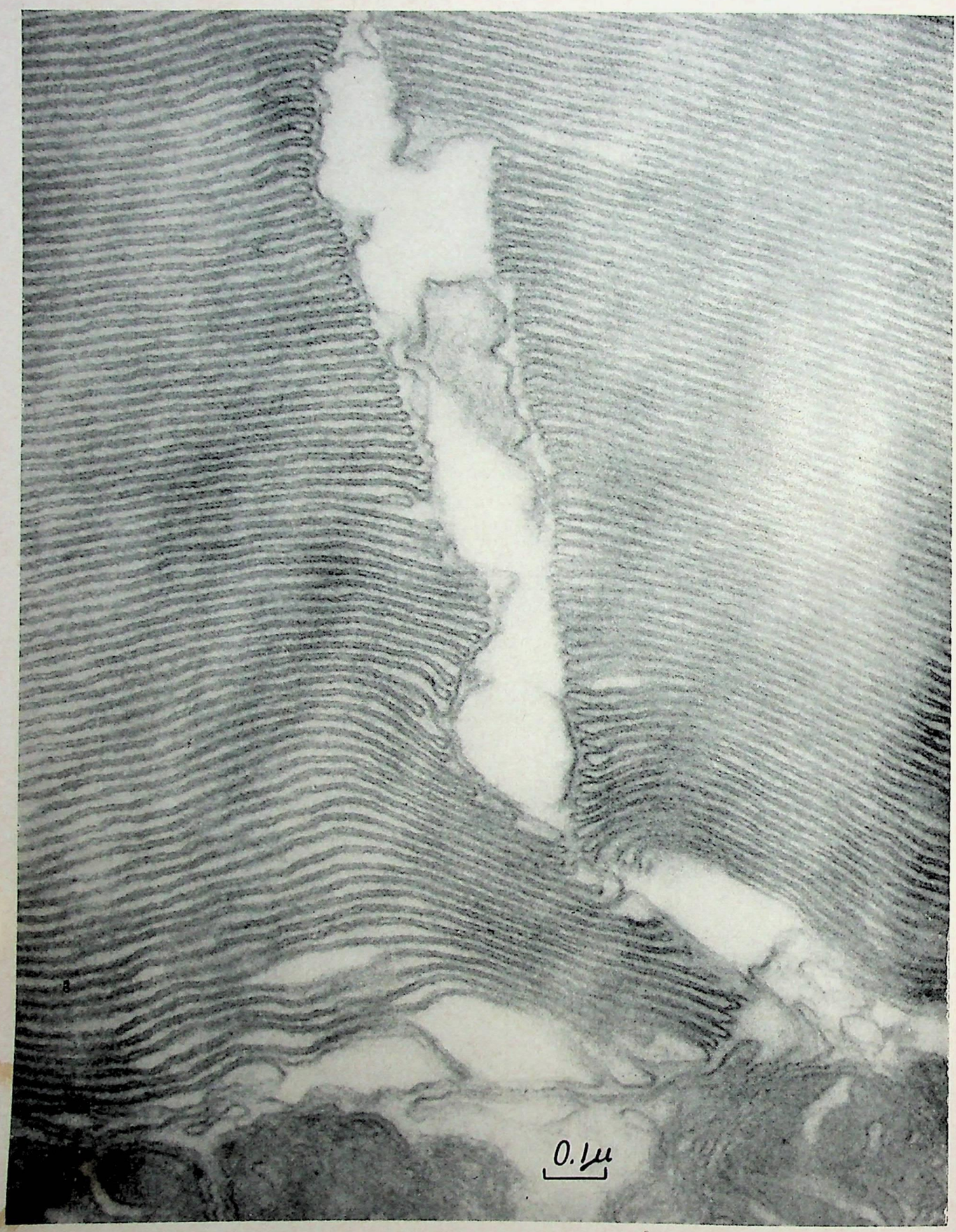
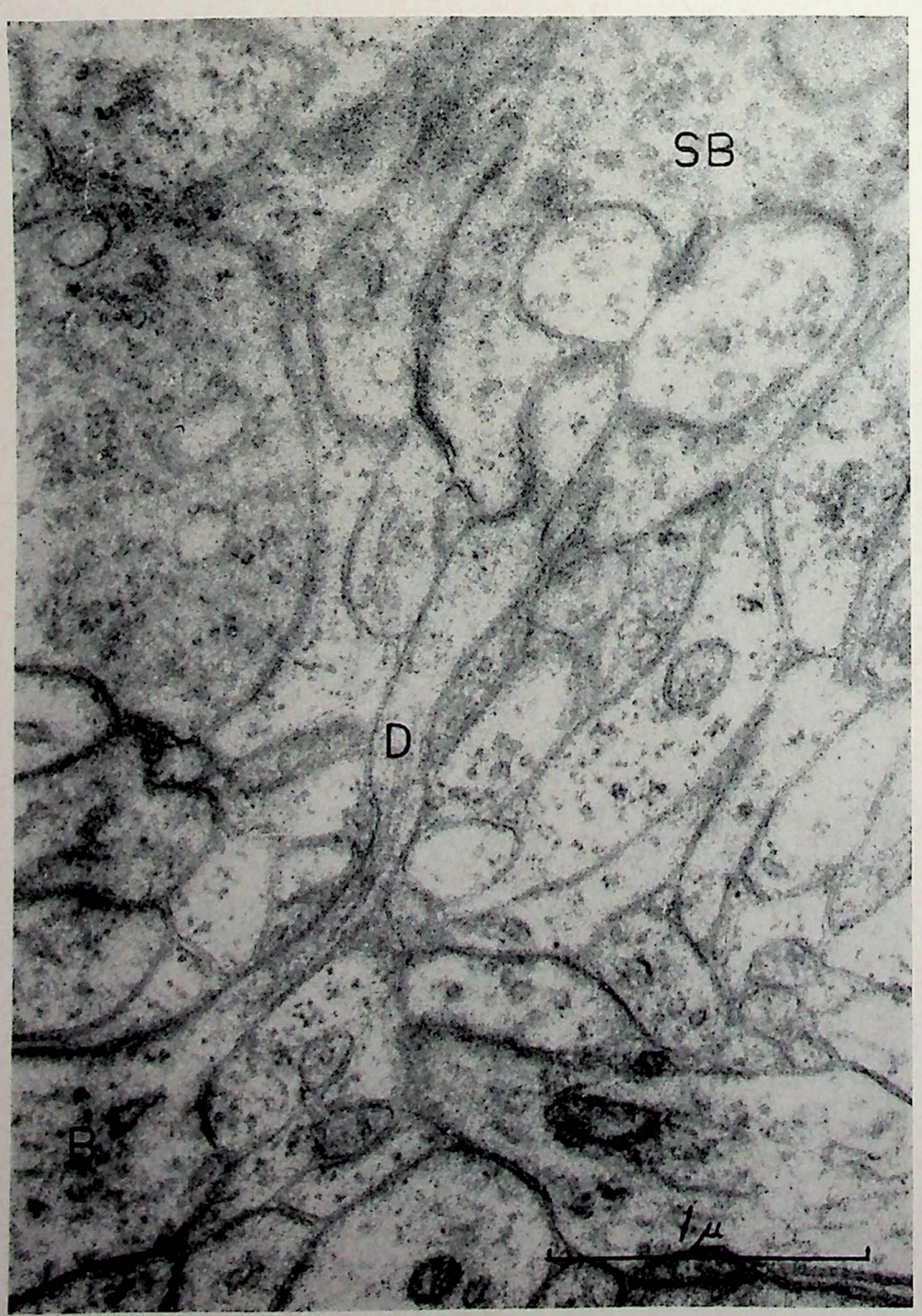


FIG. 17. The peripheral parts of two outer segments belonging to two twin cones of the perch retina. The outer segment shown in the lower half of the picture demonstrates that the less opaque interspaces in the triple-layered disks are open, and that the opaque layers of adjacent disks are continuous. At the opposite side of this segment, the conditions are those observed in the outer segment shown in the upper half of the picture. The less opaque interspaces are closed. This shows that the pile of disks is formed by a continuous membrane that is repeatedly folded. In other types of receptors, no such obvious folding can be observed but the disks are mutually connected through tube-like stalks which are continuous with the opaque layers of the disks. ×100 000.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA



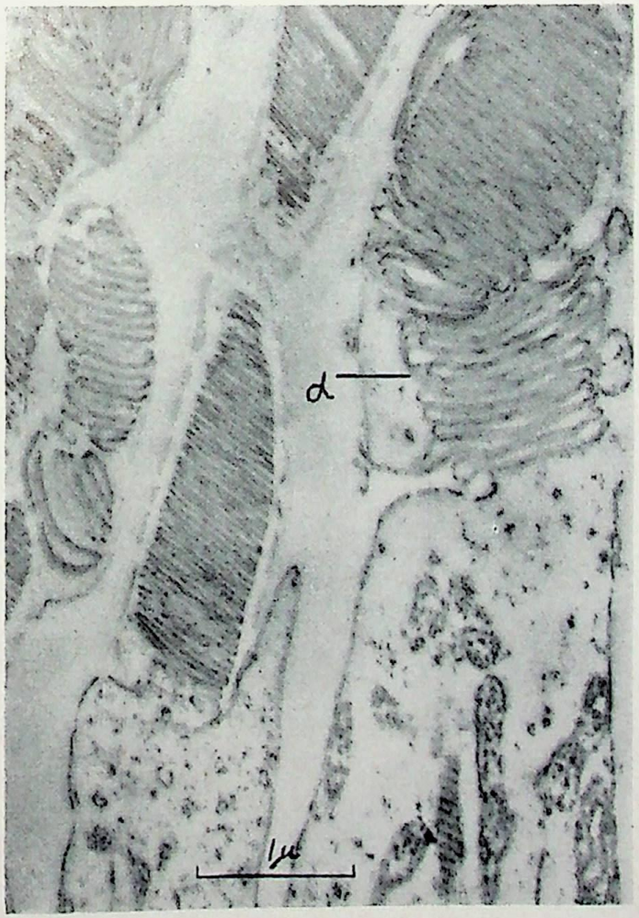


FIG. 18. Retinal receptors from an eye of a four-day-old kitten. At α, a repeatedly folded membrane is observed which appears to represent a stage of the development of the outer segment. In the cat, the retinal receptors are differentiated after the birth. ×18 000.

FIG. 19. Synaptic contact between a synaptic body, *SB*, of a retinal rod of the guinea pig eye, and a bipolar nerve cell, *B*, through the dendrite, *D*. The dendrite enters the synaptic body in an invagination of the plasma membrane and makes contact with two vacuoles also located in the same invagination [from F. S. Sjöstrand, J. Ultrastructure Research **2**, 122 (1958)].×41 000.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

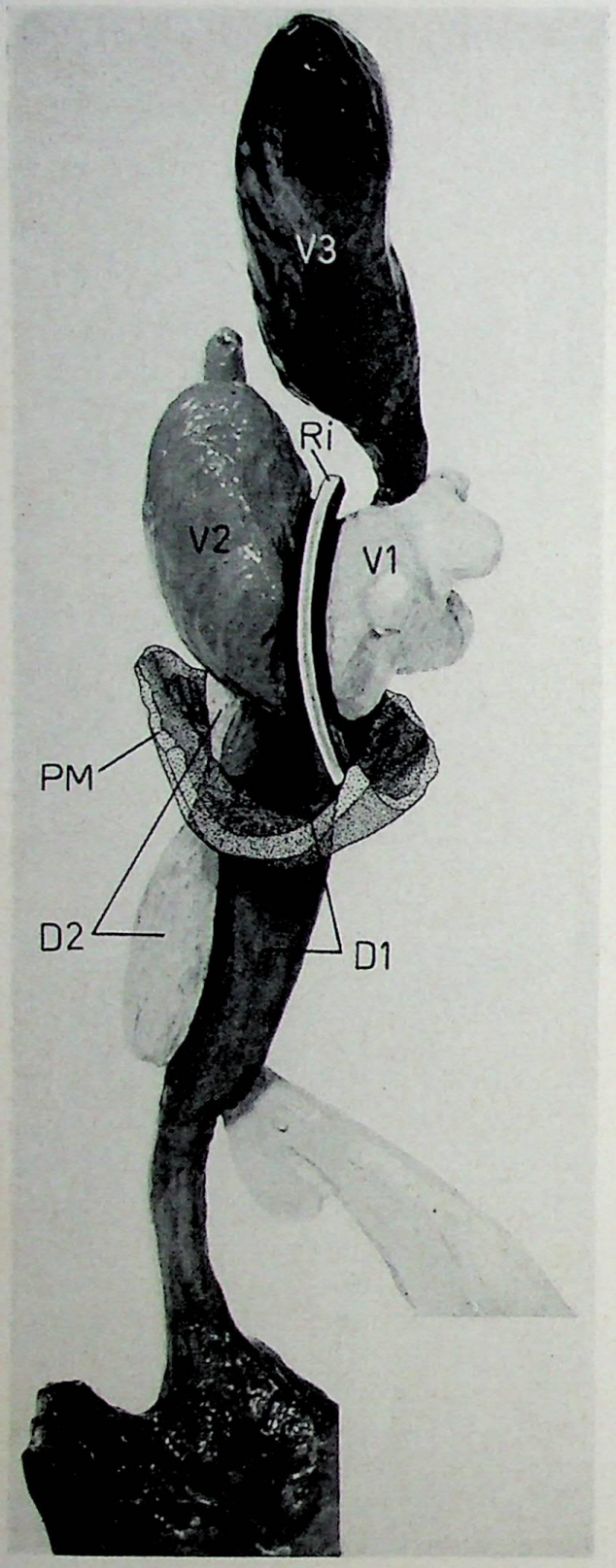


FIG. 20

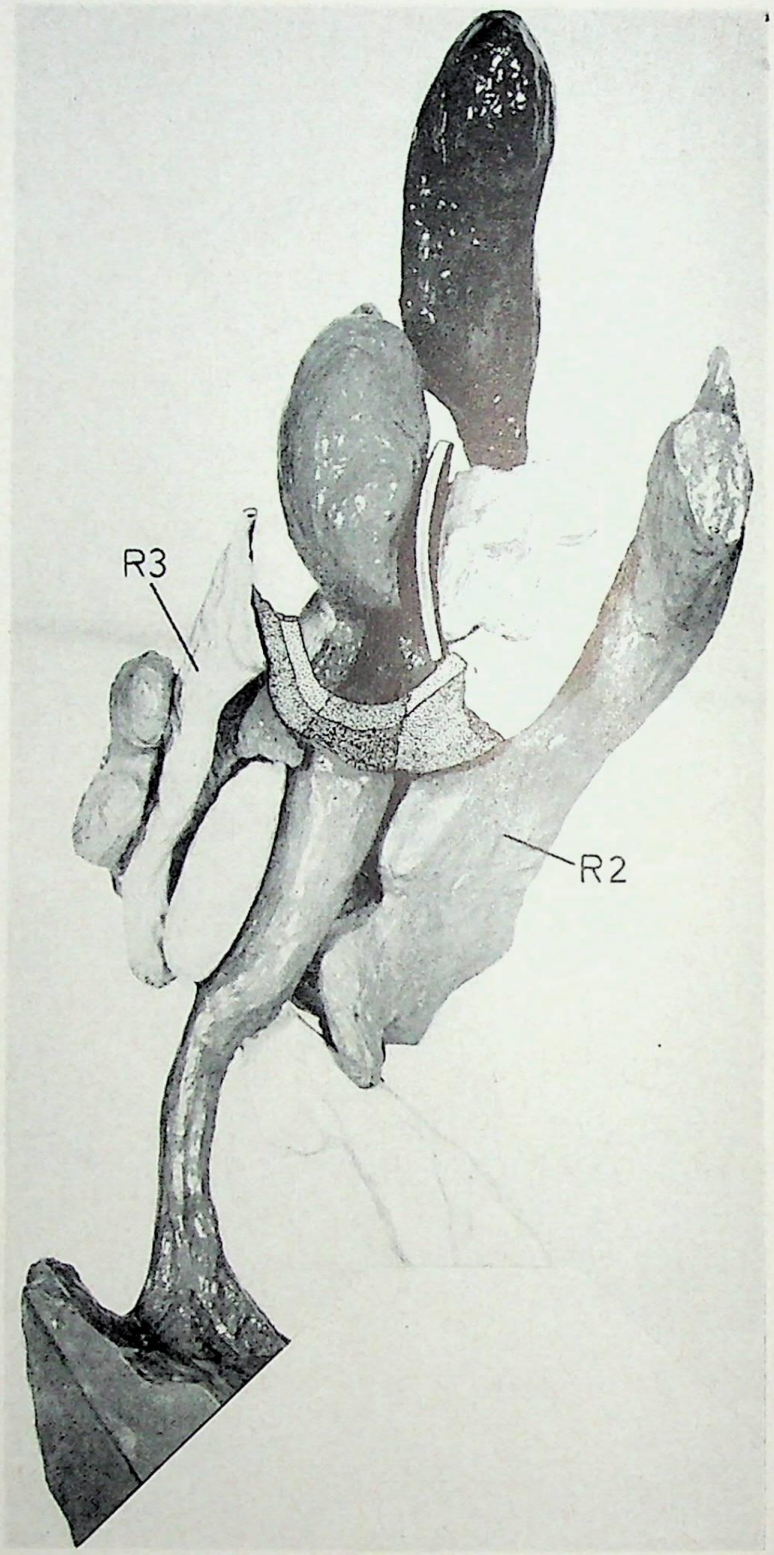


FIG. 21

FIGS. 20–21. Three-dimensional reconstruction of the synaptic body of a retinal rod of the guinea pig eye made from a series of 40 sections according to the classical technique of the three-dimensional reconstructions for series of sections. Two dendrites, D1, and D2, from nerve cells are shown entering into synaptic contact with the synaptic body of the receptor. The plasma membrane bounding the synaptic body has been removed with the exception of a fragment, PM, in order to show the large vacuoles and the synaptic ribbon which, together with synaptic vesicles and granules, constitute the elementary components of the retinal synaptic bodies. In Fig. 21, some of the extensions from adjacent synaptic bodies, R2 and R3, of receptor cells are shown making contact with the surface membrane of the reconstructed synaptic body [from F. S. Sjöstrand, J. Ultrastructure Research 2, 122 (1958)].

the outer segments appear as folds of one continuous membrane (Fig. 17). If one looks at the outer segments of the kitten retina a few days after birth, one can see the outer segments in the process of being formed. They seem to be formed through a folding of a membrane (Fig. 18), each fold being transformed into a disk later on, each disk remaining associated with its neighbors through a tube-like connection located close to the center of the disk at the end of a deep incision.

It well might be that the mitochondrial membranes

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

are formed in the same way, which explains the fact that, in 1 to 10%, the light interspace of the inner mitochondrial membranes is continuous with the light interspace of the outer mitochondrial membrane. This would give one explanation of why this kind of contact between outer and inner membranes is so rare.

Recently, we succeeded in obtaining quite extensive series of serial sections. From one series of 40 sections with an average thickness of 250 A, the author has made a three-dimensional reconstruction of the synaptic connections and the inner structure of the synaptic bodies of the retinal rods in the guinea pig eye.[28] This reconstruction has made it possible to reveal the very complicated arrangement of various membranous components of this part of the cell (Figs. 19–21) and to draw the first primitive circuit diagram (Fig. 22) of the synaptic contacts of a receptor cell. This technique is, of course, of interest when analyzing the central nervous system, but it is necessary also for a detailed description of the topographical relations between different structural components of the cytoplasm.

One now may survey some of the types of membranes as observed in the cytoplasm (Fig. 23). Distinction can be made between the cytoplasmic membranes which, after osmium fixation, are associated with the RNA particles, the infolding of the plasma membrane, and the Golgi membranes. These various types of membranes easily can be distinguished morphologically and a neutral terminology is proposed for them: α-, β-, and γ-cytomembranes.[29]

As assumed by Porter and Palade and accepted by most American electron microscopists, all of these types of membranes, as well as the nuclear membrane, the centrosphere, and various types of vesicles in the cytoplasm, have been considered to represent different aspects of one continuous canalicular system, the so-

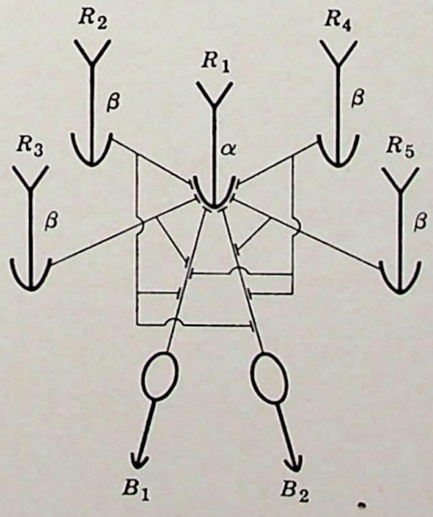


FIG. 22. A trial to make a circuit diagram of the synaptic contacts of a retinal receptor in the guinea pig eye showing the extensive inter-receptor contacts. R_1–R_5, receptor cells; B_1 and B_2, bipolar nerve cells [from F. S. Sjöstrand, J. Ultrastructure Research 2, 122 (1958)].

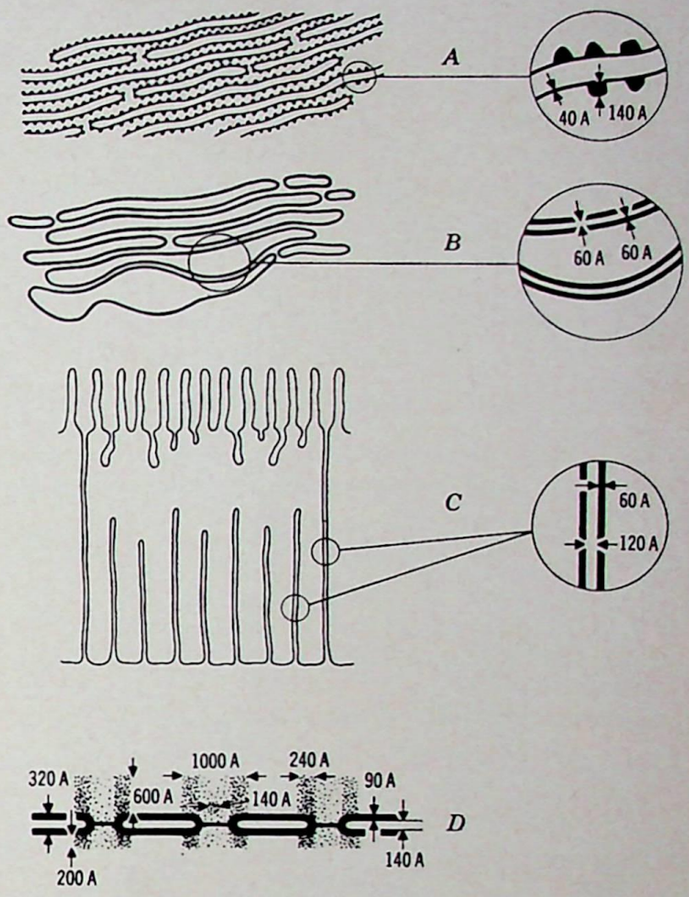


FIG. 23. Schematic presentation of some of the types of membranes that appear in cells. A. α-cytomembranes; B. Golgi membranes (γ-cytomembranes); C. β-cytomembranes; D. nuclear membrane [according to B. A. Afxelius, Exptl. Cell Research 8, 147 (1955)]. Some characteristic dimensions of the membranes are presented [from F. S. Sjöstrand, *Methods in Enzymology* (Academic Press, Inc., New York, 1957), Vol. IV, p. 391)].

called endoplasmic reticulum, extending throughout the whole cytoplasm. There are no evidences that we have been able to confirm that might allow the conclusion that these various components form a continuous system. We represent an opposite standpoint in accepting morphological differences as indicating differences regarding function, and we stress that the cytoplasm is differentiated into a limited number of structurally and presumably functionally different components.

For the future, it is probably less important to extend one's knowledge of the structure of cells to more and more types of cells. The structural patterns are repeated in a rather monotonous way. It is more important by far to try to analyze further the molecular architecture of the various components of the cytoplasm by using a variety of techniques. What appears most important, however, is to try to supplement the collected geometrical data with biochemical data. It is necessary then to work on fractions of homogenized cells, to improve the techniques of fractionation, to work out methods that make it possible to identify in the fractions the various cell components, and to estimate their relative concentrations in a quantitative way.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

Membranes appear as the most common structural principle so far. When dealing with the cytoplasm, one gets the impression that the construction of a surface membrane with certain properties represents one of the first steps in the creation of life, and that this principle has been applied in a rather monotonous way by nature.

BIBLIOGRAPHY

[1] F. S. Sjöstrand, Nature **171**, 30 (1953).

[2] F. S. Sjöstrand, J. Appl. Phys. **24**, 117 (1953).

[3] F. S. Sjöstrand and J. Rhodin, Exptl. Cell Research **4**, 426 (1953).

[4] F. S. Sjöstrand and J. Rhodin, J. Appl. Phys. **24**, 116 (1953).

[5] G. E. Palade, Anat. Record **114**, 427 (1952).

[6] G. E. Palade in *Enzymes: Units of Biological Structure and Function*, O. H. Gaebler, editor (Academic Press, Inc., New York, 1956), p. 185.

[7] F. S. Sjöstrand, J. Cellular Comp. Physiol. **33**, 383 (1949).

[8] F. S. Sjöstrand, Intern. Rev. Cytol. **5**, 455 (1956).

[9] F. S. Sjöstrand, J. Appl. Phys. **24**, 117 (1953).

[10] F. S. Sjöstrand, Experientia **9**, 68 (1953).

[11] R. S. Bear, K. J. Palmer, and F. O. Schmitt, J. Cellular Comp. Physiol. **17**, 335 (1941).

[12] F. S. Sjöstrand and T. Gierer, unpublished.

[13] W. J. Schmidt, Kolloid-Z. **85**, 137 (1938).

[14] F. S. Sjöstrand, Acta Anatomica, Suppl. 1 (1944).

[15] F. S. Sjöstrand in *Proceedings of the Conference on Electron Microscopy*, Delft, 1949, p. 144.

[16] F. S. Sjöstrand, Nature **165**, 482 (1950).

[17] H. Fernández-Morán and J. B. Finean, J. Biophys. Biochem. Cytol. **3**, 725 (1957).

[18] F. S. Sjöstrand and V. Hanzon, Exptl. Cell Research **7**, 393 (1954).

[19] K. R. Porter, J. Exptl. Med. **97**, 727 (1953).

[20] G. E. Palade and P. Siekevitz, Federation Proc. **14**, 262 (1955).

[21] F. S. Sjöstrand and R. F. Baker, J. Ultrastructure Research **1**, 239 (1958).

[22] A. J. Dalton and M. D. Felix, Am. J. Anat. **94**, 171 (1954).

[23] F. S. Sjöstrand and V. Hanzon, Experientia **10**, 367 (1954).

[24] F. S. Sjöstrand and V. Hanzon, Exptl. Cell Research **7**, 415 (1954).

[25] Å. Holmberg, *Ultrastructural changes in the ciliary epithelium following inhibition of secretion of aqueous humour in the rabbit eye.* Thesis, Stockholm (1957).

[26] H. Zetterqvist, *The ultrastructural organization of the columnar absorbing cells of the mouse jejunum.* Thesis, Stockholm (1956).

[27] J. D. Robertson, J. Biophys. Biochem. Cytol. **3**, 1043 (1957).

[28] F. S. Sjöstrand, J. Ultrastructure Research **2**, 122 (1958).

[29] F. S. Sjöstrand, *Physical Techniques in Biological Research* (Academic Press, Inc., New York, 1956), Vol. III, p. 241.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

38

Fine Structure of Biological Lamellar Systems*

H. FERNÁNDEZ-MORÁN

Mixter Laboratories for Electron Microscopy, Neurosurgical Service, Massachusetts General Hospital, Boston 14, Massachusetts, and Department of Biology, Massachusetts Institute of Technology, Cambridge 39, Massachusetts

INTRODUCTION

LAMELLAR structures represent one of the most important and widespread components of biological systems.[1,2] Ranging from the individual membranes of cells to the complex multilayered structures found in the nerve myelin sheath, photoreceptors, and chloroplasts, they embody similar patterns of molecular organization. As derivatives of cell membranes, the lamellar structures exhibit the characteristic permeability properties, electrical activity, and other functional features essential to life. Moreover, by virtue of their repetitive and orderly arrangement, the multilayered systems are uniquely endowed for performing the functions of conversion, transfer, and storage of energy in organisms.

Investigation of the morphological substrate of these processes has been handicapped by the marked lability of membranous systems when subjected to the effects of analytical techniques. However, with recent advances, a general picture of the organization of these systems at the macromolecular level is now gradually emerging. The successful application of high-resolution electron microscopy in combination with polarization-optical and x-ray diffraction studies has revealed the extraordinary degree of regularity and structural differentiation which is present throughout the various hierarchies of organization down to the molecular level.

This review describes the salient features of the fine structure of the nerve myelin sheath, photoreceptors, and other representative lamellar systems. It is becoming increasingly evident that all lamellar systems appear to have certain common structural parameters at the molecular level, possibly reflecting an underlying functional analogy. Therefore, as more information becomes available on any one of these specialized systems, new approaches are suggested in the correlation of structure and function. Thus, many of the biochemical studies and recent concepts derived from solid-state physics in investigating the mechanisms of photosynthesis,[3-7] may eventually prove to be of operational value in elucidating the primary events of nerve function and sensory reception.

FINE STRUCTURE OF THE NERVE MYELIN SHEATH

The myelin sheath of nerve fibers is one of the most highly ordered biological systems, and exhibits all of the properties of the smectic fluid-crystalline state. Despite the high water content and marked lability of the myelin sheath, its ultrastructure can be studied by combined application of different techniques under various conditions. The analysis of the fine structure of the myelin sheath can be regarded, in fact, as one of the best examples of the systematic application of complementary biophysical and biochemical methods. Since the myelin sheath derives from a multiply-folded Schwann-cell surface,[8,9] it can also be considered a model system for the study of cell-membrane structure in general.

Polarization-optical studies of the myelin sheath[10-12] indicated that it is composed of concentrically arranged protein or lipoprotein lamellae which alternate with layers of lipid molecules oriented with their long axes in the radial direction. This concept was confirmed and extended by x-ray diffraction studies of fresh nerve.[13-18] It was assumed that the thickness of the concentric lipid-protein unit layers corresponded to the fundamental small-angle x-ray diffraction spacings of 170 A recorded in amphibian, and of 180 to 185 A found in mammalian peripheral nerves.

Subsequently, the postulated layers and the concentric laminated structure of the sheath were observed directly in electron micrographs of thin nerve sections fixed with osmium tetroxide.[19-23] However, since the preparation of these specimens required fixation, dehydration and embedding in a plastic medium, followed by slicing into ultrathin sections, the possibility of artifacts had to be seriously considered. Moreover, the regular patterns of selective deposition of osmium could not be interpreted in terms of specific regions containing the lipids, lipoproteins, water, and other components of the normal myelin structure.

One is obviously dealing here with a case in which combined application of low-angle x-ray diffraction techniques and high-resolution electron microscopy can supplement each other in many ways. It is instructive to consider the merits and drawbacks of each technique in order to appreciate the type of approach required in analyzing biological systems.

The x-ray diffraction method offers the great advantage of permitting examination of the intact nerve trunk in the living animal.[13,15] The information derived from the x-ray diffraction pattern of normal active nerve can, therefore, be considered a reliable basis for all structural analyses of the myelin sheath.[13,18,24] From the low-angle x-ray diffraction pattern, the dimensions and

* Part of these studies were aided by a research grant (C-3174) from the National Institute of Neurological Diseases and Blindness of the National Institutes of Health, Public Health Service, U. S. Department of Health, Education, and Welfare.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

approximate distribution of scattering groups in the radial direction of the sheath can be deduced. Reliable reference parameters are, therefore, provided for checking the results of electron microscopy in studies of the normal and modified myelin sheaths.[18,24]

Finally, the x-ray diffraction pattern recorded in a few hours represents an average of the main structural parameters of all of the nerve fibers contained in the exposed nerve trunk which contribute to the diffraction. The practical significance of this is immediately apparent when comparing it with the far greater time factor involved in the corresponding electron-microscope study. Examination of nearly 100 000 serial ultrathin cross sections would be required for a complete electron-microscope study of a 1-mm long segment of a few fibers only.

Conversely, electron microscopy offers the unique advantage of revealing directly the complex patterns of macromolecular organization in selected areas of the specimen.

Structure of the Normal Myelin Sheath

X-ray diffraction patterns recorded from fresh peripheral nerve in a direction perpendicular to the fiber axes feature a series of very well-defined reflections at low angles. The wide-angle reflections at about 4.7 A exhibit meridional intensifications, while the low-angle reflections are precisely oriented in the equatorial direction. The low-angle reflections can be interpreted in terms of a fundamental radial repeating unit varying from 170 A in amphibians to 184 A in mammalian peripheral nerve. The low-angle reflections also show a characteristic alternation of intensities in the even and odd orders [Fig. 1(b)]. It is assumed that the fundamental repeating unit consists of two parts having very similar distributions of x-ray scattering power; from the intensities of the odd-order reflections, the magnitude of the difference between the two parts can be estimated.[18] Finean[15–18] has determined that this "difference factor" is appreciable in peripheral nerve myelin, but appears to be negligible in the optic nerves.

High-resolution electron micrographs of ultrathin sections of osmium-fixed peripheral nerve (Fig. 1) show a series of concentrically arranged, dense lines separated by lighter spaces with an average period of 130 to 140 A. There is, thus, a correspondence of the periodic layers observed in the electron micrographs of myelin preparations and the fundamental radial unit derived from low-angle x-ray diffraction patterns.

The discrepancy of 20 to 30 A between the layer thickness and the x-ray fundamental spacing is owing to the shrinkage effects introduced by osmium fixation and other preparative procedures connected with the examination of the thin sections in the electron microscope.[24] In addition to the periodic dense bands about 30 A wide, the electron micrographs feature light intermediate bands which are much narrower (10 to 15 A). This variation in the densities of the two principal bands of osmium deposition is related to the "difference factor" which manifests itself in the low-angle pattern as an increase in the intensity of the first-order reflection.

The present lack of precise information on the chemistry and localization of the sheath proteins, the lipids, the water layers, and other components of myelin has not as yet permitted a direct correlation with the structural data.[25–27] However, by combined electron-microscope and x-ray diffraction studies of controlled physical and chemical modifications of the myelin sheath, a general picture of the molecular arrangements in the radial units can be obtained.[15,18,21,24,28] Satisfactory correlation and interpretation of structural relationships at the macromolecular level requires the preparation of a large number of ultrathin (100 to 200 A) undistorted serial sections of the same specimens utilized for the parallel x-ray diffraction studies. This has been greatly facilitated by improved thin-sectioning techniques using a diamond knife.[24,29]

Such a combined approach has led to a detailed analysis of the preparation procedures, thus validating important findings and defining artefact sources. The extraction and enzymatic-digestion experiments performed on fresh myelin proved to be particularly revealing.

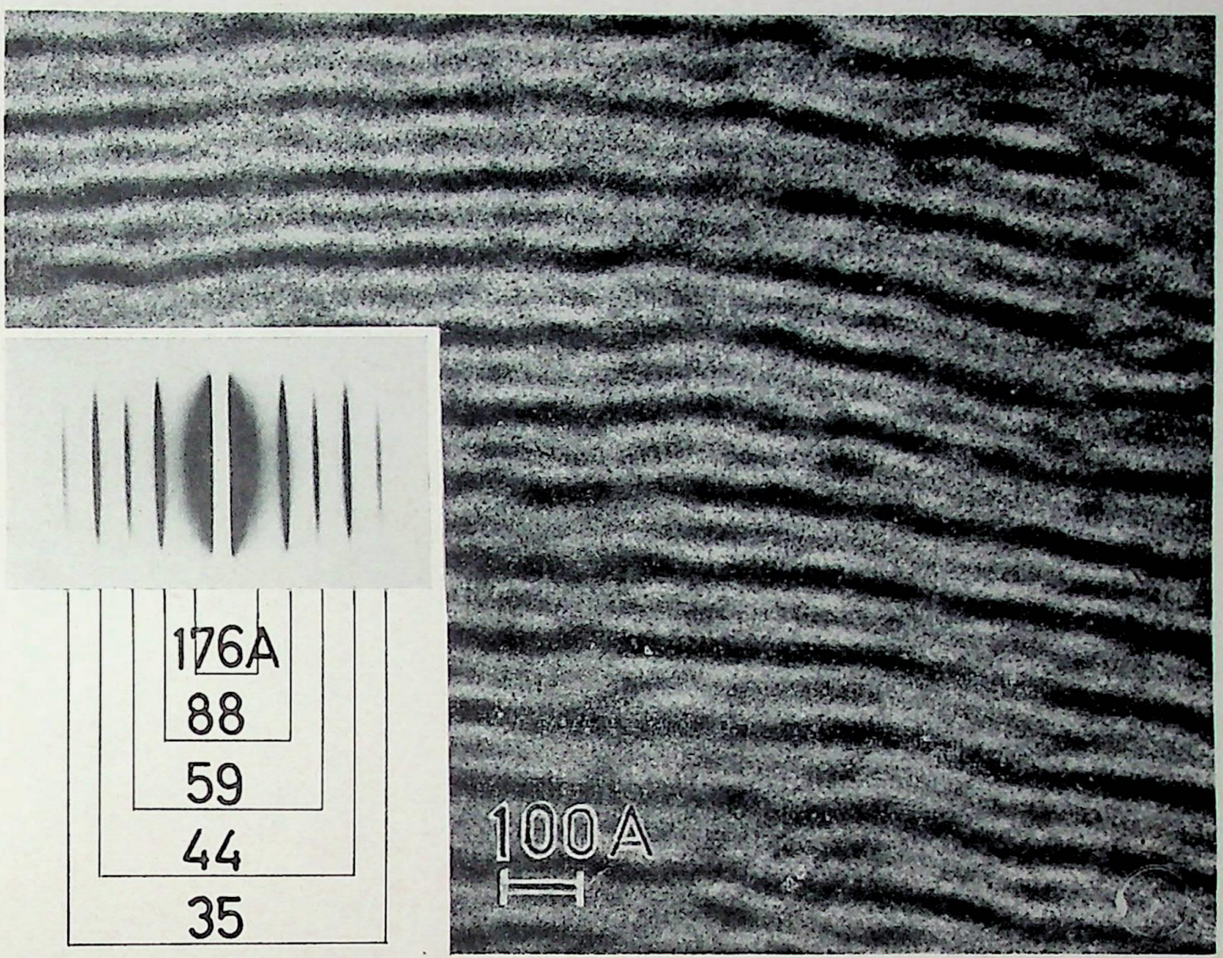
Lipid-Extraction Experiments

Extraction of fresh nerve with acetone at 0°C removes about 30% of the cholesterol,[18] leaving the other lipid components essentially intact within the framework of a still highly organized residual myelin sheath.[24] The main modifications revealed by electron microscopy and x-ray diffraction [Figs. 2(d) and 2(e)] of the residual sheath comprise expansion of the layered structure with internal rearrangements and formation of collapsed layer systems [Figs. 2(a) and 2(b)]. The dense lines are relatively resistant to lipid solvents, and their persistence in the collapsed layered structure may indicate that the heaviest osmium deposition is in the region of the protein or stable lipoprotein constituents. These modifications emphasize the importance of cholesterol in the myelin structure, and lend further support to the existence of the phospholipid-cholesterol complex postulated in the unit cell by Finean.[18]

Low-angle diffraction patterns of the lipid extract show a strong 34.2-A reflection [Fig. 2(c)] characteristic of cholesterol. When the dried lipid extract is examined by electron microscopy, thin crystalline lamellae are found embedded between condensed multilayered aggregates. These crystalline lamellae give electron-diffraction patterns which are similar to those recorded from pure cholesterol.

A more extensive breakdown of the sheath is observed after alcohol extraction. The fine-layered structures observed in certain areas may represent the refractory protein framework in addition to recrystallized lipid components. The extraction experiments suggest, therefore, that the dense bands observed in electron micro-

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA



(b) (a)

FIG. 1. (a) High-resolution electron micrograph of myelin-sheath segment from a transverse section of an osmium-fixed rat sciatic nerve embedded in gelatin. The average layer spacing is 110 A. Note compact and well-preserved dense lines with moderate enhancement of the intermediate lines. ×850 000. (b) Low-angle *x*-ray diffraction pattern of rat sciatic nerve recorded with Finean camera. This pattern features a fundamental period of 176 A, with characteristic alternation of the intensities of the even and odd orders.

graphs of the normal myelin sheath represent areas of selective osmium deposition at lipoprotein interfaces, whereas the light bands are regions occupied by lipid chains and associated myelin components which do not react primarily with osmium tetroxide.

Enzymatic-Digestion Experiments

Although digestion with trypsin is not specific, its application to fresh nerve fibers produces characteristic modifications of the myelin fine structure.[19,24] In addition to slight expansion of the concentric layers, the uniform dense lines appear to dissociate into rod-shaped granules, 30 to 50 A wide and 40 to 60 A long (Fig. 3). The unit myelin lamellae isolated from the sheath after trypsin digestion likewise exhibit extensive dissociation into composite elongated granules of similar dimensions.[19,20] An analogous granular fine structure of the dense layers is commonly found after freezing and thawing of fresh nerve, $KMnO_4$ fixation, and other types of preparations.[24] Moreover, there are also numerous indications of a compact granular fine structure in the dense layers of the normal myelin sheath.

X-ray diffraction data furnish supporting evidence for the presence of a regular organization within the plane of the layers. A strong 60- to 70-A vector, which appears necessary to account for the relative intensities of the low-angle reflections, can be related to this type of organization within the planes of the lipoprotein layers.[24]

Nerve-Degeneration Studies

The breakdown of myelin known to occur during *in vitro* degeneration of nerve affords an opportunity to study the fine structure of the disintegrating sheath without introducing extraneous reagents. Comprehensive studies of nerve-fiber degeneration, which are now being carried out by the author with combined application of high-resolution electron microscopy and low-angle x-ray diffraction techniques, reveal interesting details of the layer structure.

When fresh peripheral nerve is enclosed in a glass capillary under aseptic conditions and left to degenerate at 20°C, low-angle x-ray diffraction patterns recorded at periodic intervals show characteristic changes

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

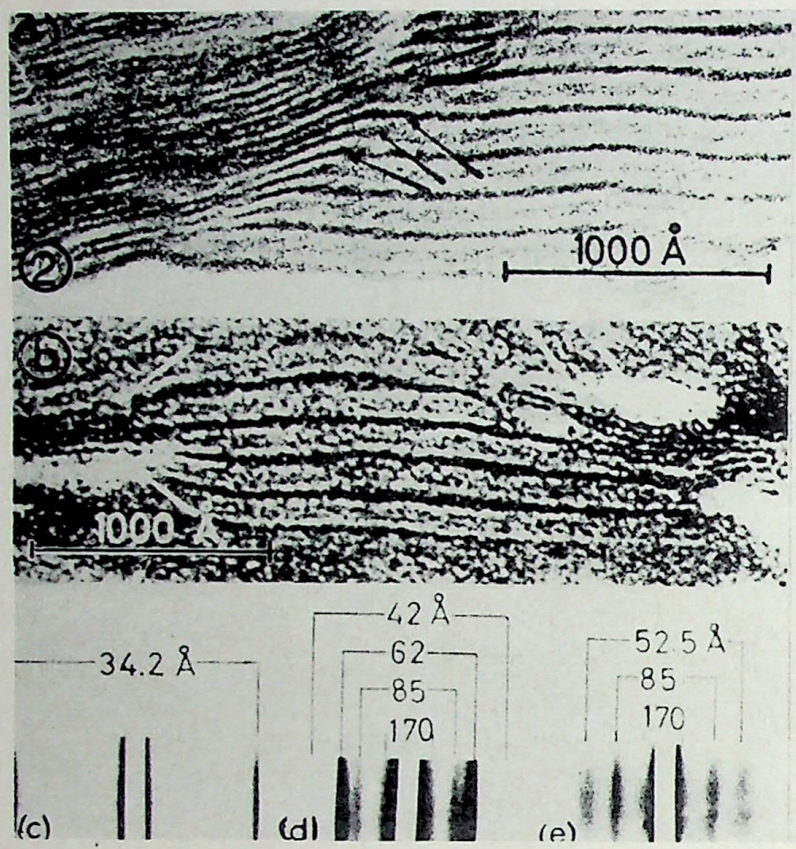


FIG. 2.

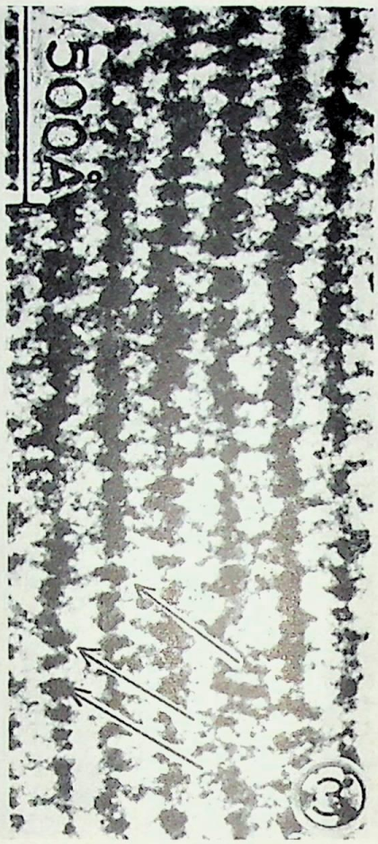


FIG. 3

FIG. 2. (a) Myelin-sheath segment from transverse section of fresh rat sciatic nerve extracted with acetone at 0°C for 12 hours prior to fixation with 2% osmium tetroxide and embedding in butyl methacrylate. Notice the expanded periods (160 A) of the modified layers at the right, and the transitions (arrows) to the collapsed period (43 A) at the left. ×300 000. (b) Myelin-sheath segment from longitudinal section of fresh rat sciatic nerve extracted with acetone at 0°C for 12 hr prior to fixation with osmium tetroxide and embedding in butyl methacrylate. Notice the transitions (arrows) from the expanded layer system to the collapsed layer system ×280 000. (c) Low-angle x-ray diffraction pattern of the lipid material extracted from fresh sciatic nerve by immersion in acetone at 0°C for 12 hours, showing the characteristic cholesterol spacing at 34.2 A. (d) Low-angle x-ray diffraction pattern of residual dried rat sciatic nerve after extraction with acetone at 0°C for 12 hours. (e) Low-angle x-ray diffraction pattern of acetone extracted nerve fixed with buffered 2% osmium tetroxide and embedded in methacrylate.

FIG. 3. Myelin-sheath segment from a rat sciatic nerve which was incubated with 1% crystalline trypsin for 12 hours prior to fixation in buffered 1% isotonic osmium-tetroxide solution and embedding in butyl methacrylate. Notice the marked dissociation of the dense lines (arrows) into elongate granules. ×520 000.

(Fig. 4). Compared with the normal pattern, marked modifications of the distribution of x-ray scattering power within the unit are noted as degeneration progresses. There is marked intensification of the second-order diffraction, followed by the appearance of a 70-A reflection, and gradual extinction of the lower orders in later stages of degeneration. The corresponding electron micrographs (Fig. 5) show various forms of granular dissociation of the dense layers, resembling the structures observed after trypsin digestion. In addition to this supporting evidence of granular structure within the layers, numerous other features of myelin organization are being uncovered as the complex dissolution process is followed at the submicroscopic level.

Suggested Molecular Organization of Myelin

Based largely on their pioneering x-ray diffraction studies, Schmitt and co-workers[12–14,30] concluded that the fundamental radial unit of the myelin sheath in the internodal portion of peripheral nerve comprised two lipoprotein layers, each of which consisted of bimolecular leaflets of 67 A sandwiched between protein layers of 25 A, with interposed water layers contributing a further 25 A. Taking into consideration the contraction of the lipid layers during drying, and the results of other experimental modifications, Finean[15–18] arrived at a more detailed picture of the molecular arrangement in the myelin unit. According to this conception, each lipoprotein layer would consist of two phospholipid-cholesterol complexes associated with a cerebroside molecule, intercalated between monolayers of protein. The two units are distinguished by a "difference factor" which has not been accurately determined, but can nevertheless be explained by the mechanism of myelin formation. As Geren[8,9] showed, the myelin sheath is formed by an infolding and multiple wrapping of the Schwann-cell membrane around the axon in embryonic fibers. Assuming that the Schwann-cell lipoprotein membrane is asymmetric, then this process of rolling

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

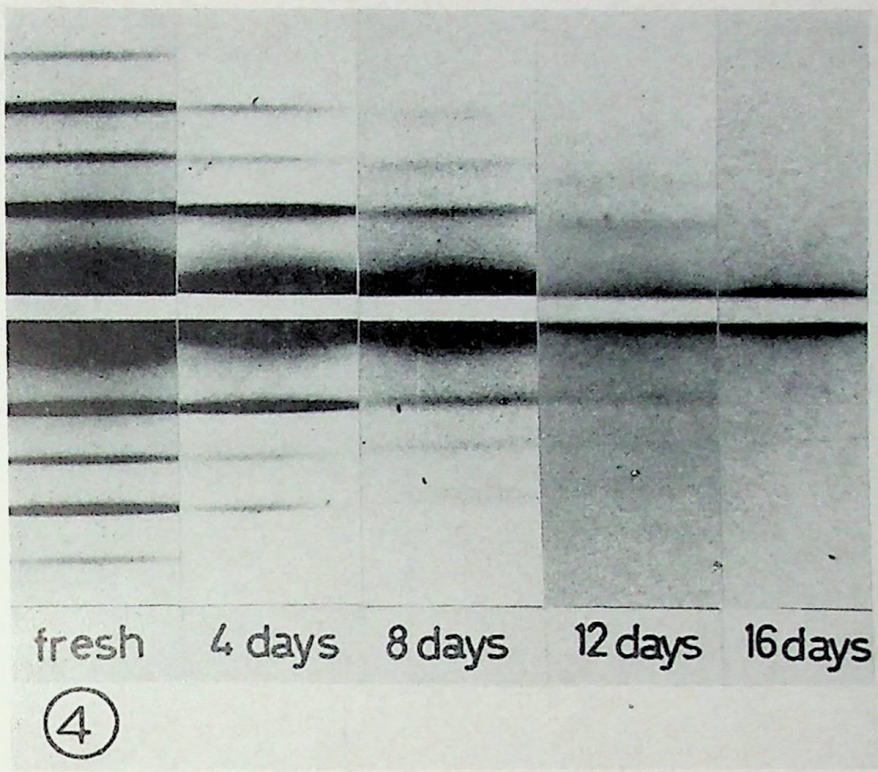


FIG. 4.

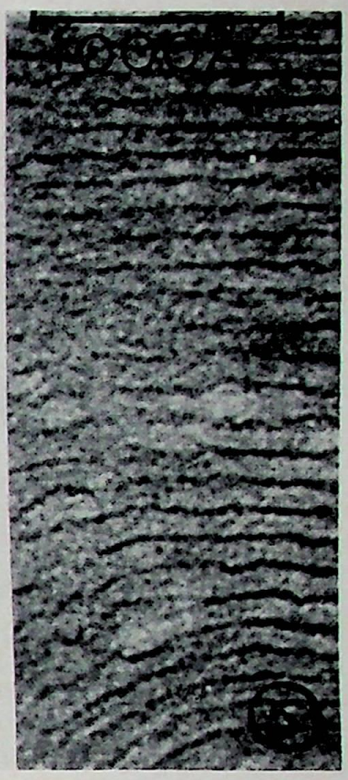


FIG. 5.

FIG. 4. Low-angle x-ray diffraction patterns of rat sciatic nerve recorded during *in vitro* degeneration at 20°C. Notice marked modifications of the distribution of x-ray scattering power within the radial unit as degeneration progresses.

FIG. 5. Electron micrograph of nerve myelin sheath during *in vitro* degeneration (4 days), showing dissociation of the dense lines into granular structures, disappearance of the intermediate line, and other modifications of the layered structure. ×280 000.

onto the axon will produce the observed symmetry difference in successive layers.[18]

The lipids are considered to be oriented with their long axes radially in the sheath. The 4.7-A, meridionally intensified ring recorded from fresh nerve has been related to the cross sections of the lipid molecules, indicating the average interchain separation of their hydrocarbon chains. However, in order to account for the permeability properties of the membranes, various possibilities including the existence of real or potential "holes" have been discussed. One of the suggested configurations envisages a continuous layer of lipids sandwiched between monolayers of protein with a fenestrated or open-network type of structure. The filtering action of the protein layer would combine with the selective solubility of the penetrating ions or molecules in the lipid layer to account for the sieve mechanism of permeability. The formation of submicroscopic pores by radial extensions of the protein layers across the lipid layers has also been considered, but no direct evidence for the existence of these discontinuities is yet available. Nevertheless, the x-ray diffraction and electron-microscope data clearly indicate that there must be a considerable degree of organization within the plane of the lipoprotein layers which remains to be investigated.

Water is one of the most important components of fresh myelin, constituting at least 35% of the myelin sheath.[12,30] Information on the precise localization of water within the myelin layers is also of great importance when considering the possible pathways of diffusion of ions and other solutes between the axon and the extracellular fluids.[28,31] The water layers at the aqueous interfaces of the fundamental repeating unit in the compact myelin are about 12 to 15 A thick.[31] Under certain experimental conditions,[17] larger amounts of water can be incorporated between the aqueous interfaces in the myelin unit. The water is presumably coordinated on the protein layer located at the aqueous surface of each bimolecular leaflet. The thickness of the water layers is largely determined by the electrical charge density at the aqueous interfaces,[31] and by the ionic strength, since water is expelled from lipid layers in the presence of ions.[14] These changes in thickness of the aqueous interfaces may be of significance in connection with ion movements in the intermembrane spaces. However, as Schmitt[31] has pointed out, the physicochemical properties, such as *p*H and ionic strength in capillary spaces of these macromolecular dimensions, would be quite different from those in bulk solutions. The effects of the electrically charged lipid groups on the structure of the hydration water in these aqueous interfaces may also induce the formation of a crystalline arrangement resembling "frozen" hydration sheaths postulated in protein molecules.[32]

Through the application of high-resolution nuclear

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

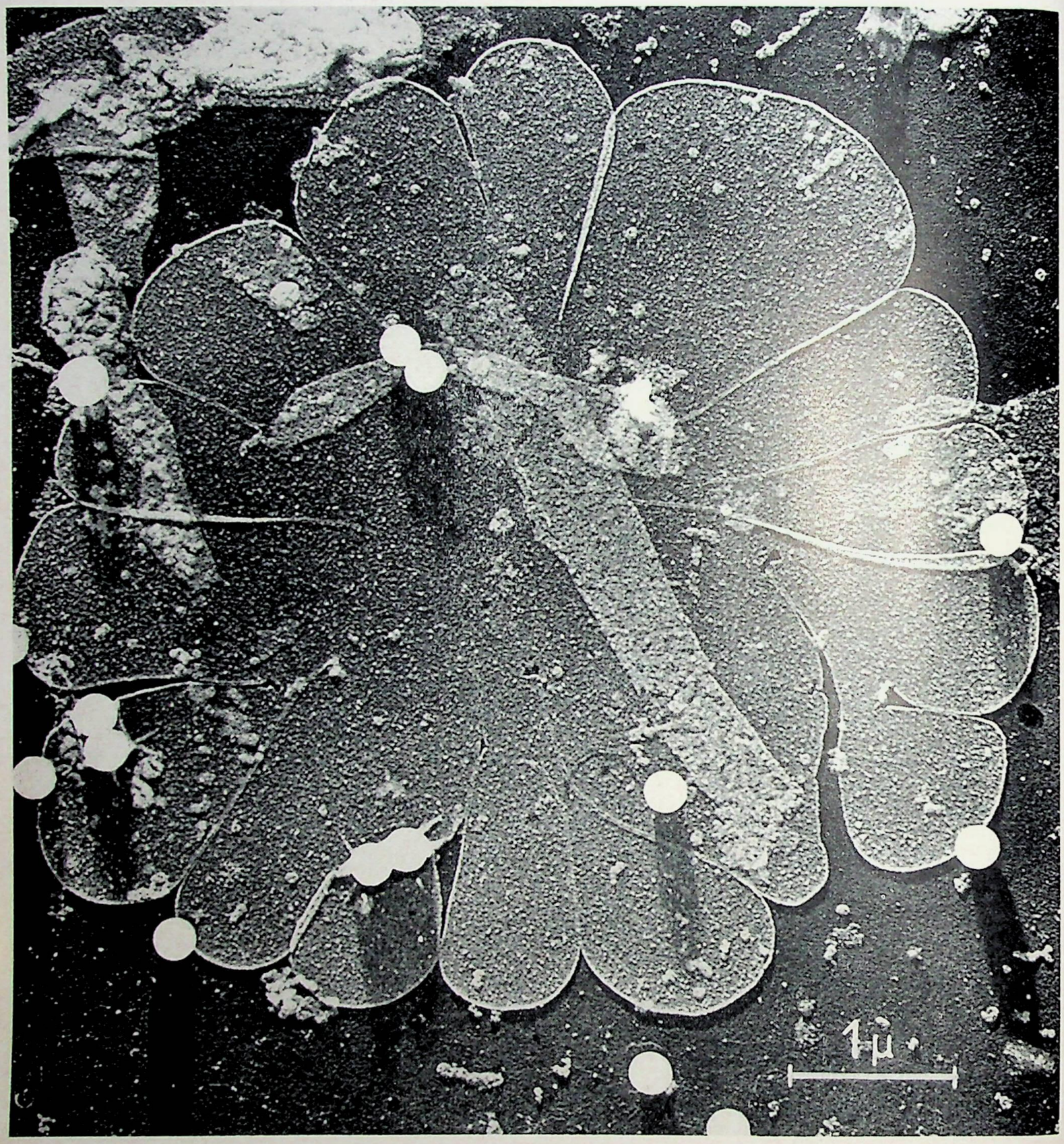


FIG. 6. Single unit disk isolated from the outer segment of frog retinal rod, showing the granular surface structure and the marginal cord. Latex particles (2800 A diam) have been added for calibration in this shadowed preparation. ×25 000.

magnetic-resonance spectrometry, it now appears possible to obtain important supplementary information on the water content of fresh whole nerve and the hydration state of its components, in a rapid and nondestructive way.[33] The size of the proton magnetic-resonance signal recorded from biological materials is proportional to the water content.[34] Since the mobility of the protons in the solid, nonaqueous constituents is much less than in the aqueous phase, the resonance spectrum consists of a narrow line owing to the water, superimposed on a broad line owing to protons in the supporting solid. After suitable calibration, the water content can be accurately determined by evaluating the line width of the proton resonance, or by measuring the amplitude of the derivative of the narrow absorption curve. Both methods have been applied in preliminary

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

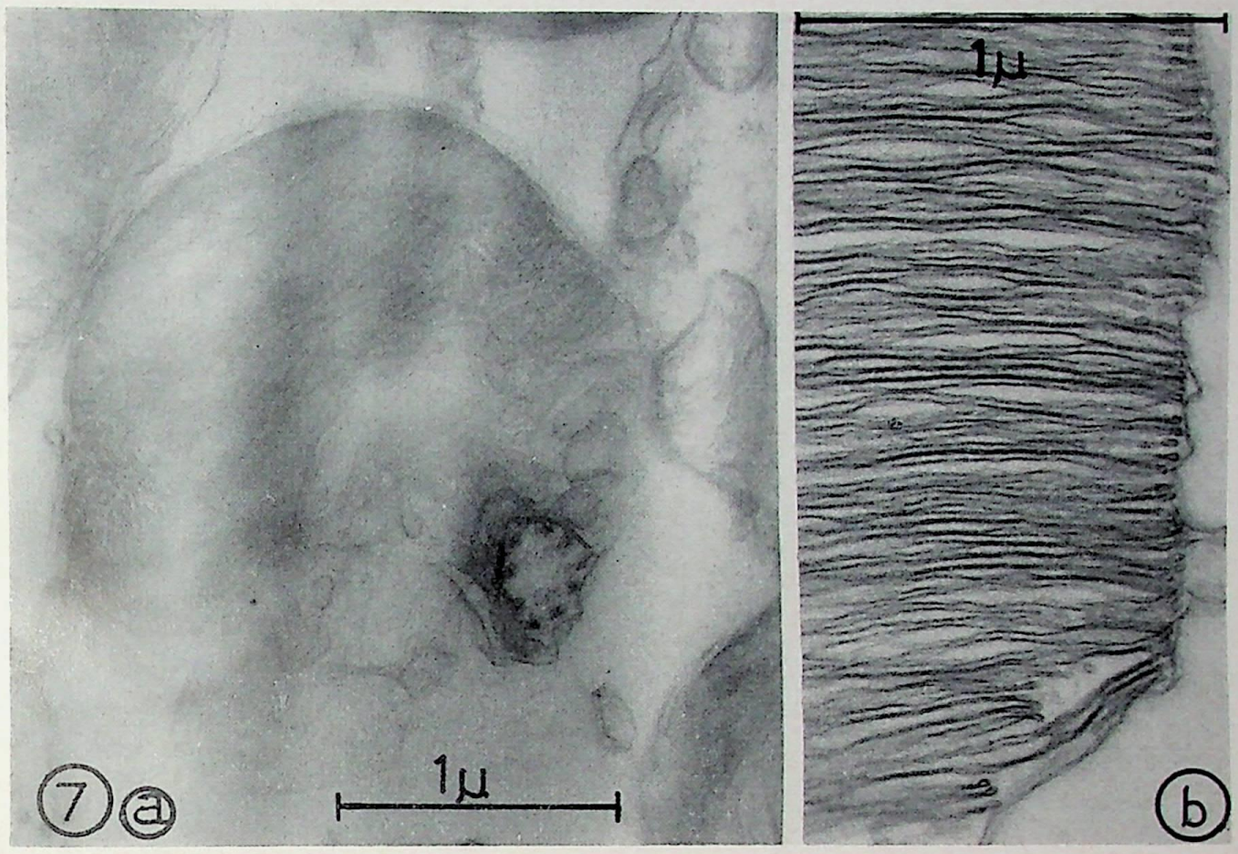


FIG. 7. (a) Transverse section through the basal region of the outer segment of a retinal rod of the guinea pig, showing the characteristic tubular structures and vesicular formations. Notice cross-sectioned filaments which establish connection with the inner rod segment. ×32 000. (b) Longitudinal ultrathin section through the basal part of the outer segment of retinal rod from guinea-pig eye showing the regular arrangement of the double membrane disks. ×48 000.

determinations of the water content of fresh nerves under various conditions, using commercially available equipment and a special transistorized NMR spectrometer.[33,35]

The protons in the hydration shells postulated around the macromolecules lack the high mobility of the water in the bulk aqueous environment, and give a broad and weak resonance line. By comparison with the strong narrow line owing to the more-mobile water protons, the proportion of water forming the hydration shells of nucleic acids has been estimated.[36] Similar experiments can be carried out on living nervous tissues, and should yield valuable information on the hydration state of the bound water in the myelin sheath.

The present concepts of the myelin's molecular architecture refer almost exclusively to the lipoprotein framework which stands up to various analytical techniques. Within this framework, no detailed localization is yet possible of the numerous enzymes, electrolytes, trace metals, and other essential components of the myelin sheath. Although deuterium and radiophosphate studies[37] indicate that myelin has a relatively low metabolic turnover, more information is required on the structural relationships with the phospholipases, phosphatases, and other enzymes involved in lipid metabolism.[38,39] Likewise, the molecular models will be incomplete until specific localizations can be assigned to the cholinesterases[40] and related enzymes which play an important role in nerve conduction. Recent investigations which link the metabolism of copper with certain demyelinating diseases[38] also focus attention on the possible relationships of trace metals with the lipoprotein layers. In view of the close association of the porphyrins[41] and related photodynamic agents with the central myelin, it may prove of interest to establish a more specific correlation with the ultrastructure of the myelin sheath.

FINE STRUCTURE OF PHOTORECEPTORS

Electron microscope studies have revealed that most of the photoreceptors in biological systems, where light is converted into chemical or electrical energy, feature a regular lamellar fine structure. In the vertebrate eye, the photoreceptor elements are composed of submicroscopic plates orderly disposed in a pile; whereas in the invertebrates, closely packed thin-walled tubules build up the light percipient constituents. The process of visual excitation is initiated by the absorption of light in the characteristic photopigments which form an integral part of all photoreceptors. In the vertebrates, the visual pigments, rhodopsin and iodopsin, are located in the outer segments of the rods and cones, respectively. Polarized-light studies[42] had already indicated that the rod outer segments consist of transversally oriented protein layers alternating with layers of longi-

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

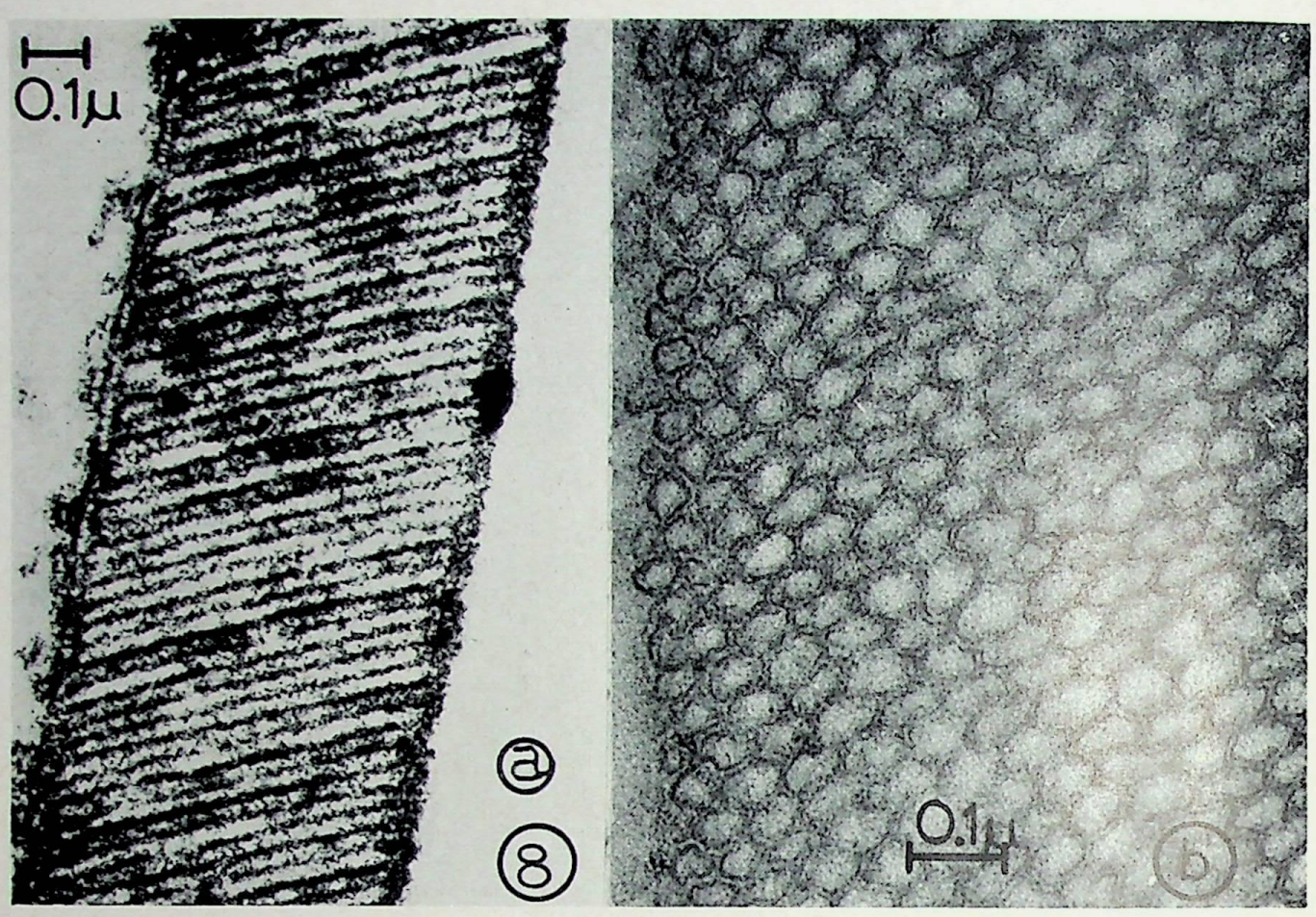


FIG. 8. (a) Oblique longitudinal section through a rhabdomere of the housefly retinula showing the lamellar type of internal structure. The highly regular submicroscopic organization is illustrated by the nearly perfect parallel array of several score dense bands occupying the entire rhabdomere. These parallel bands correspond to the dense walls of the tubular compartments sectioned longitudinally. ×60 000. (b) High-resolution electron micrograph of longitudinal ultrathin section through a rhabdomere of the housefly retinula, demonstrating the fine structure of the tubular compartments. The regular array of sharply-outlined annular profiles (400 to 500 A in diameter) bounded by a dense osmiophilic substance is interpreted as representing cross sections of the closely packed tubular compartments. This differentiation of the dense "walls" of the compartments into a thin, osmiophilic boundary line, 20 to 30 A wide, associated with a dense layer of approximately 60 A, may be of interest in relation to the postulated models of the macromolecular photoreceptor complexes. ×100 000.

tudinally oriented lipid molecules, and could, therefore, be compared with a radial cylinder of the myelin sheath. Electron microscopy confirmed that the entire rod outer segment consists of several hundred unit disks, about 150 A thick, stacked up in regular array.[21,43,44]

The unit disks of the frog retinal rods have a diameter of about 6μ and exhibit a peculiar lobulated shape with numerous incisures outlined by a dense marginal cord (Fig. 6). Thin sections disclose the remarkably regular layering of the unit disks [Fig. 7(b)], which are arranged with their planes perpendicular to the rod axis. Each disk consists of dense double membranes which stain intensely with osmium and enclose a light space of about 70 to 80 A. The repeating unit along the axis of the rod in osmium-fixed preparations corresponds closely to the spacing of about 320 A recorded in the low-angle x-ray diffraction patterns.[45] Recent studies demonstrate tubular processes at the incisures which are closely associated in the basal portion with the bundle of thin fibrils connecting the outer and inner rod segments [Fig. 7(a)]. The inner segments contain numerous densely-aggregated long mitochondria,[44,46] which are regularly encountered in all types of photoreceptors.[47,48] The retinal cones show a similar layered structure.[44]

It has been tentatively assumed that the protein membranes stain densely with osmium, while the lipid molecules may be predominantly localized in the intermembrane compartments of the disk.[44] However, knowledge of the detailed chemical composition of the photoreceptors is still too meagre, and controlled extraction and digestion studies of the type described earlier remain to be done. The exact location of the visual pigments is also unknown, but Wolken and co-workers[49] have suggested that the pigment molecules are oriented as monolayers at the aqueous protein and lipoprotein interfaces. The chloroplasts exhibit a very similar layered structure, and the chloroplastin-pigment complex appears likewise to reside in the aqueous protein-lipoprotein interfaces.

Light Receptors in Insect Compound Eyes

In the insect compound eye, each separate visual element or ommatidium consists of eight sensory retinula cells.[50] The differentiated medial part of each retinula cell, known as a rhabdomere, is built up of approximately 20 000 to 30 000 closely packed tubular

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

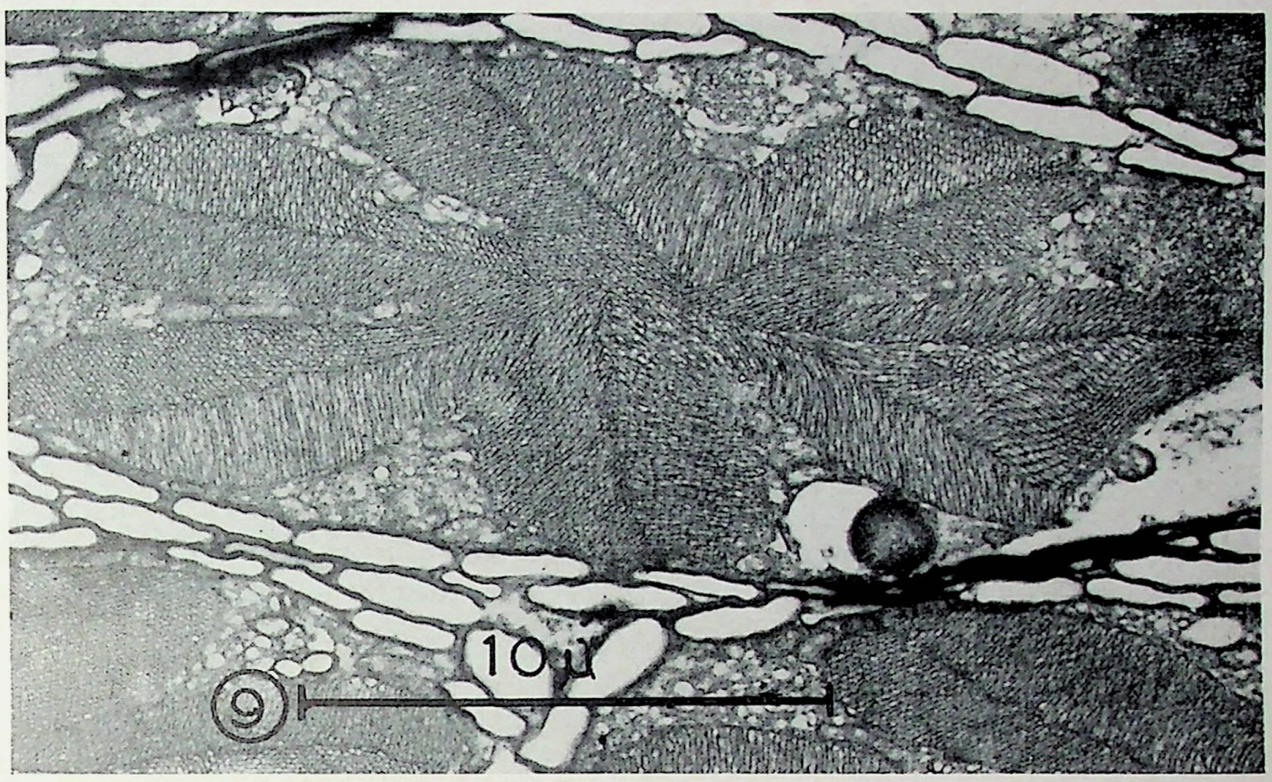


FIG. 9. Oblique cross section through a retinula from the eye of the giant tropical moth, *Erebus*, demonstrating the symmetrical arrangement of the fused rhabdomeres. Note that the general configuration and orientation of the tubular compartments of any one of the rhabdomeres is matched by the corresponding rhabdomere pattern directly or diagonally opposite to it. The equivalent rhabdomeres of contiguous ommatidia are also oriented with their main axes in the same direction, thus introducing a degree of "structural polarization" in the over-all pattern which must bear relation to the analysis of polarized light in the insect eye. ×6000.

compartments which vary from 400 to 1200 A in diameter, and are oriented in highly regular array [Fig. 8(a)] with their long axes normal to the longitudinal rhabdomere axis.[47–49,51] These tubular rhabdomere compartments are regarded as differentiated microvilli of the retinula cell membrane. The walls of the tubular compartments exhibit a dense boundary line of about 30 A in which the oriented visual pigments may possibly be located[47] [Fig. 8(b)]. Within each ommatidium, the rhabdomeres are radially arranged in a symmetrical pattern formed by matched pairs of opposite rhabdomeres displaying similar orientation of their layered structure (Fig. 9).

This differentiated submicroscopic organization of the visual elements—and their structural coupling in coordinated groups[47]—suggests a correlation with the remarkable ability of insects to recognize the regional patterns of polarization of the sky as a basis for light-compass orientation.[52–54] It is assumed that the radially arranged rhabdomere pairs containing dichroic visual pigments similarly oriented in their periodic tubular compartments may correspond to the functional units of the analyzer for polarized light postulated in the compound eye.[47–49,51,55]

The visual system of the insects and other arthropods features an intimate association with a complex three-dimensional network of air-filled tubules or tracheoles and with pigment granules, which is probably of considerable functional significance.[47]

The structural analysis of the insect compound eye is proving to be one of the most revealing examples of how the highly differentiated tubular and lamellar textures can provide a periodic array of gas-liquid-solid interfaces at the supramolecular level, which are ideally suited for selective interaction (absorption, reflection, diffraction, scattering, etc.) with incoming light signals.

RELATION OF LAMELLAR STRUCTURES TO ENERGY-TRANSFER PROCESSES

Essentially the same type of macromolecular architecture described in the unit layers of the myelin sheath and the photoreceptors is encountered with only minor modifications in the organization of the mitochondria, intracytoplasmic lamellar complexes, Golgi apparatus, and other specialized lamellar systems.[21,31,33] If this striking structural similarity is actually the expression of an underlying functional analogy, then interesting general correlations may be derived from critical collation of the available information.

Correlation of fine structure with energy-conversion processes has been most fruitful in those systems—like mitochondria and chloroplasts—which can be isolated in quantity and studied outside of the cell without appreciably impairing their functional activity. Biochemical studies[56–58] have shown that certain fragments of mitochondria are still capable of carrying out the essential functions of electron transport and oxidative phosphorylation. Based on this evidence, the mitochon-

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

drion has been compared with a giant polymer made up of certain repeating units incorporated in the internal and external membranes.[57] Recently, it has been possible by digitonin treatment to obtain fragments of the mitochondrial membrane which contain relatively intact "respiratory chain assemblies" including the enzymes necessary for coupling phosphorylation to electron transport.[58] These fragments contain protein, phospholipid, and a small amount of cholesterol;[58] they may possibly correspond in many respects to the double membrane components obtained by sonic fragmentation of mitochondria.[56,57] The available evidence indicates that these enzymes and related components of the electron-transfer chain are precisely arranged in specific patterns built into the matrix of the lipoprotein layers, which provide the highly ordered "floor space" where these assemblies are anchored.[57,58]

It has been suggested that the electrons are transported in this comparatively rigid structure, approximating the solid state, by being relayed through small, mobile molecules within the lipoprotein matrix.[56,57] Considering the important participation of water in all lamellar systems, it is tempting to associate the water molecules or hydrated small molecules with this transfer function.

The general problem of structural interaction between an ordered substrate and its aqueous environment appears to be of particular significance in the fluid-crystalline lamellar systems with their high water content. Many unique features of proteins in aqueous solutions suggest that the macromolecules may be surrounded by a hydration sheath of icelike character.[32,36,59] The induction of a crystalline lattice of hydration water is ascribed to the local and long-range cooperative electric-field effects of the nonpolar side chains in proteins.[32] In the case of the 10- to 15-A thick, aqueous layers intercalated between the protein-lipoprotein layers, these effects are expected to be even more pronounced in determining the formation of crystalline lamellae of "frozen" water.

On first consideration, this concept of "frozen" hydration lamellae permeating in periodic array such a multilayered complex would seem to be incompatible with the smectic fluid-crystalline character of biological systems. However, it should be borne in mind that the "icelike" hydration sheaths lodged in ultracapillary spaces of these molecular dimensions may exhibit physical and physicochemical properties which are quite different from those encountered in bulk solutions. Also, only that part of the hydration water which is in immediate contact with the protein or lipoprotein layers may be ordered in a crystalline lattice; while the remaining "liquid" aqueous strata, and the additional interstitial water channels would still allow sufficient freedom of movement of the flexible layered substrate. In this connection, it is interesting to note that the adsorbed hydration water in graphite facilitates slippage of the crystalline lamellae, and is largely responsible for the lubrication properties of this material.

It is, therefore, conceivable that the postulated crystalline hydration sheaths could provide an organized framework interconnecting the lipoprotein layers and serving primarily as a transfer medium for charged carriers like electrons, holes, protons, and ions. By virtue of the extensive hydrogen bonding in the icelike lattice, the effects of ion movements could be produced over comparatively long distances without involving bodily displacement of the ions, which would be an important factor in the relatively compact lamellar systems.

Regardless of these hypothetical assumptions, it is obvious that much more must be known about the hydration state and properties of water in lamellar systems before detailed functional correlations can be attempted. Nuclear magnetic-resonance spectrometry, supplemented by spin-echo techniques, now offers the possibility of obtaining essential quantitative data on relaxation times and correlation times of the water components in living biological systems under various physiological conditions. It is also suggested that a structural determination of the hydration shells of lattice-ordered water may eventually be possible by extending the neutron-diffraction studies already successfully carried out on ice[60] to suitable models of multilayered systems, or to liquid crystals of adequate size.

Important additional evidence that the processes of energy transfer and conversion in the chloroplasts are dependent on the highly ordered molecular arrangement in their layered structure has been furnished by comprehensive recent studies of the mechanisms of photosynthesis.[3–7,61] Calvin and co-workers[3,4,6] have strongly indicated that the initial process of energy transfer is a physical one, dependent on the specific physical organization of the chlorophyll and its associated pigments in the ordered, "quasicrystalline" layered structures. Absorption of a light quantum by chlorophyll raises an electron to a conduction band, leaving behind a "hole," both of which migrate together as an "exciton" through the oriented array of chlorophyll molecules. The conjugated carotenoid molecules act as conductors of electrons, while the lipids in the chloroplasts would function as an insulator, permitting a separation of charges when the electrons are "trapped" on one side of the layer and the "holes" on the other. This polar character is similar to that which exists at a *p-n* semiconductor junction. The electrons participate in the formation of a reducing agent leading to the reduction of carbon dioxide; the positive holes react with water to form oxygen, thus initiating the two basic reaction sequences of the photosynthetic mechanism. Direct experimental evidence has also been presented for the existence of trapped electrons and holes by recording electron spin-resonance signals from illuminated whole chloroplasts[3–6,61] at low temperatures

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

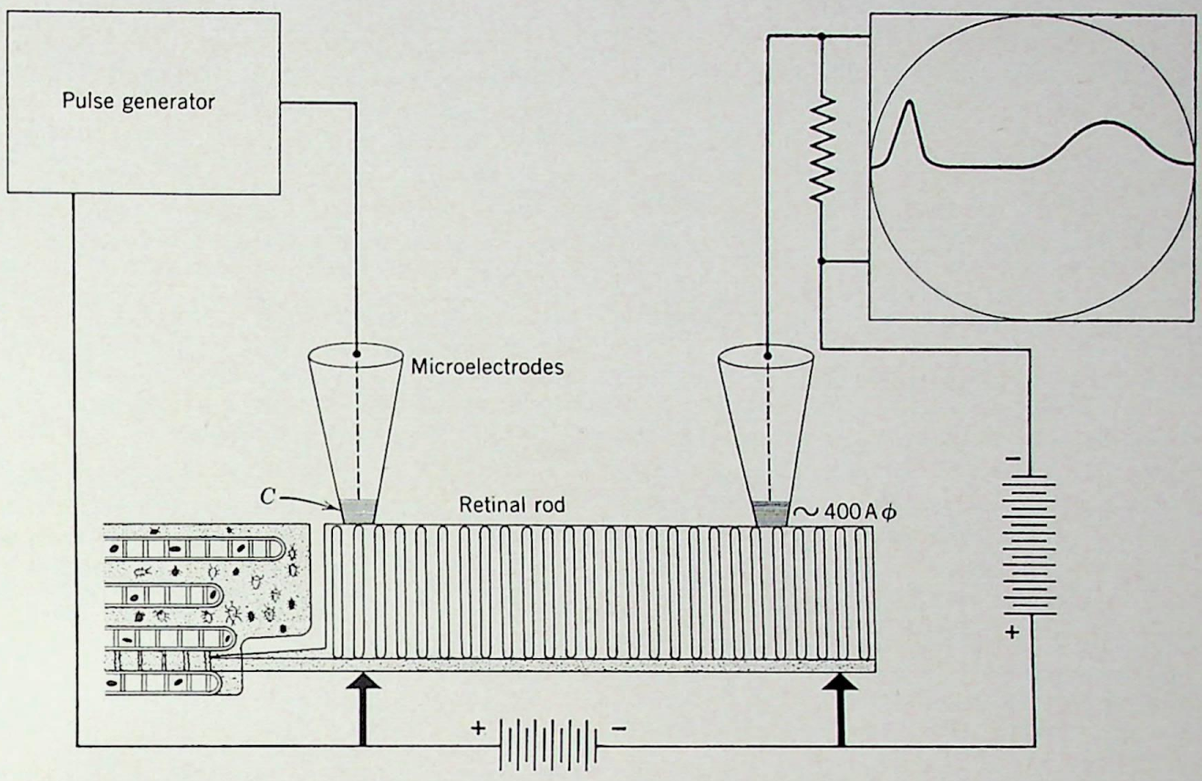


FIG. 10. Suggested experimental arrangement to investigate semiconductor phenomena associated with the mobility of electrons and holes in active biological lamellar systems. By applying microelectrodes (filled with a suitable solid free radical, "*C*") directly to the outer segments of retinal rods, it might be possible to determine the drift velocity and density of injected current carriers. Addition of a transverse magnetic field would then permit determination of the Hall effect and other magnetic concentration effects. Experiments of this type, carried out on photoreceptors and chloroplasts or on artificial multilayered structures containing specific photopigments, are needed to furnish direct experimental proof of the postulated semiconductor properties in biological systems.

(−150°C), and, therefore, excluding the possibility of any ordinary enzymatic process.

By tentatively extending these concepts to other lamellar systems, one arrives at a general picture of paracrystalline lipoprotein layers containing assemblies of specific enzymes associated with photopigments or specialized electron-transfer systems, all of which are organized in highly ordered patterns. These biological lamellar systems, permeated with hydration water of possible icelike character, might have many properties in common with semiconductors, although, as pointed out earlier,[21,62] the concepts of solid-state physics are not directly applicable to the fluid crystalline state. However, merely as a working hypothesis, the semiconductor analogy should prove useful as a possible basis for the physicochemical amplifying devices suggested in connection with the events leading from the primary process of sensory excitation to the complex bioelectric phenomena associated with nerve-impulse propagation.[63,64]

Although these ideas are admittedly speculative, they may, nevertheless, serve to encourage new experimental approaches. Of particular interest would be systematic investigations designed to furnish direct experimental proof in biological lamellar systems of the two processes of electronic conduction in semiconductors, corresponding to positive and negative mobile charges, which are characteristic of transistor action.[65] In the suggested experiment (Fig. 10), two capillary microelectrodes filled at the tip with a suitable solid free radical are applied directly to the outer segments of freshly isolated retinal rods, or, alternatively, to large chloroplasts. With the outlined arrangement, it might be possible to determine the drift velocity and the density of injected current carriers, in analogy to the classical experiments[65] carried out on small samples of germanium using micromanipulator techniques. By introducing a transverse magnetic field, determination of the Hall effect and other magnetic-concentration effects[65] should also be possible. Application of these techniques to simple model systems containing multilayers of organic semiconductors would probably be the first step in this type of study. The study of organic semiconductors[66,67] is just beginning now, and further developments in this field are bound to have far-reaching implications in the interpretation of biological phenomena.

Ultrastructural studies are confronted with more immediate tasks as a result of recent improvements in the preparation techniques and the useful resolving power of electron microscopy. Thus, continuation of earlier experiments based on rapid freezing of glycerine-treated tissues with liquid helium,[33] followed by special fixation and embedding procedures at low temperatures, may ultimately permit a more reliable correlation of organization at the molecular level with sequentially arrested states of activity in biological systems. Application of

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

moire effects[68] to directly visualize equivalent configurations of atomic dimensions may also prove to be of considerable operational value in the study of paracrystalline components.

From the foregoing cursory survey, it is evident that the correlative investigation of fine structure and function of lamellar systems poses some of the most challenging and rewarding problems in modern biology, which will require an integrated approach enlisting the best efforts of biophysicists, biochemists, and physicists.

ACKNOWLEDGMENTS

The author gratefully acknowledges stimulating discussions with M. Calvin, F. O. Schmitt, and J. Townsend, and their generous assistance in furnishing important bibliographical references.

BIBLIOGRAPHY

[1] A. Engström and J. B. Finean, *Biological Ultrastructure* (Academic Press, Inc., New York, 1958).

[2] A. Frey-Wyssling, *Submicroscopic Morphology of Protoplasm* (Elsevier Publishing Company, Inc., Amsterdam, 1953), second edition.

[3] M. Calvin and P. B. Sogo, Science **125**, 499 (1957).

[4] G. Tollin, P. B. Sogo, and M. Calvin, "Energy transfer in ordered and unordered photochemical systems," University of California Radiation Laboratory Rept. (1957).

[5] P. B. Sogo and B. M. Tolbert, Advances in Biol. and Med. Phys. **5**, 1 (1957).

[6] J. A. Bassham and M. Calvin, *Currents in Biochemical Research*, D. E. Green, editor (Interscience Publishers, Inc., New York, 1956), p. 29.

[7] C. P. Whittingham, Progr. in Biophys. and Biophys. Chem. 320 (1957).

[8] B. B. Geren and J. Raskind, Proc. Natl. Acad. Sci. U. S. **59**, 880 (1953).

[9] B. B. Geren, Exptl. Cell Research **7**, 558 (1954).

[10] W. J. Schmidt, *Die Doppelbrechung von Karyoplasma, Zytoplasma und Metaplasma* (Gebrüder Borntraeger, Berlin, 1937).

[11] F. O. Schmitt and R. S. Bear, J. Cellular Comp. Physiol. **9**, 261 (1937).

[12] F. O. Schmitt and R. S. Bear, Biol. Revs. Cambridge Phil. Soc. **14**, 27 (1939).

[13] F. O. Schmitt, R. S. Bear, and G. L. Clark, Radiology **25**, 131 (1935).

[14] F. O. Schmitt, R. S. Bear, and K. J. Palmer, J. Cellular Comp. Physiol. **18**, 31 (1941).

[15] J. B. Finean, Exptl. Cell Research **5**, 202 (1953).

[16] J. B. Finean, Expt. Cell Research **6**, 283 (1954).

[17] J. B. Finean and P. F. Millington, J. Biophys. Biochem. Cytol. **3**, 89 (1957).

[18] J. B. Finean, Exptl. Cell Research **5**, 18 (1958).

[19] H. Fernández-Morán, Exptl. Cell Research **1**, 143 (1950).

[20] H. Fernández-Morán, Exptl. Cell Research **3**, 282 (1952).

[21] H. Fernández-Morán, Progr. in Biophys. and Biophys. Chem. **4**, 112 (1954).

[22] F. O. Schmitt, Proc. Assoc. Research Nervous Mental Diseases **28**, 247 (1950).

[23] J. D. Robertson, J. Biophys. Biochem. Cytol. **1**, 271 (1955).

[24] H. Fernández-Morán and J. B. Finean, J. Biophys. Biochem. Cytol. **3**, 725 (1957).

[25] I. H. Page, *Chemistry of the Brain* (Charles C. Thomas, Springfield, 1937).

[26] J. Folch-Pi and W. M. Sperry, Ann. Rev. Biochem. **17**, 147 (1948).

[27] H. J. Deuel, Jr., *The Lipids* (Interscience Publishers, Inc., New York, 1951), Vol. I.

[28] F. O. Schmitt and N. Geschwind, Progr. in Biophys. and Biophys. Chem. **8**, 166 (1957).

[29] H. Fernández-Morán, Ind. Diamond Rev. **16**, 128 (1956).

[30] F. O. Schmitt in *Genetic Neurology*, P. Weiss, editor (The University of Chicago Press, Chicago, 1950), p. 40.

[31] F. O. Schmitt, Exptl. Cell Research, **5**, 33 (1958).

[32] I. M. Klotz, Science **128**, 815 (1958).

[33] H. Fernández-Morán in *Metabolism of the Nervous System* Second International Neurochemical Symposium, E. Richter, editor (Pergamon Press, London, 1957), p. 1.

[34] T. M. Shaw and R. H. Elsken, J. Chem. Phys. **26**, 565 (1953).

[35] P. Denis, A. Csaki, M. Delco, J. Sprenger, H. Fernández-Morán, and W. Rawyler, Arch. sci. (Geneva) **10**, 223 (1957).

[36] B. Jacobson, W. A. Anderson, and J. T. Arnold, Nature **173**, 772 (1954).

[37] W. M. Sperry and H. Waelsch, Proc. Assoc. Research Nervous Mental Diseases **28**, 255 (1950).

[38] C. E. Lumsden, *Modern Trends in Neurology* (Butterworths Scientific Publications, London, 1957), Ser. II, p. 130.

[39] G. H. Sloane-Stanley, Biochem. Soc. Symposia **8**, 44 (1952).

[40] D. Nachmansohn in *Modern Trends of Physiology and Biochemistry*, E. S. G. Barron, editor (Academic Press, Inc., New York, 1952), p. 229.

[41] H. Klüver in *Biochemistry of the Developing Nervous System*, International Neurochemical Symposium, H. Waelsch, editor (Academic Press, Inc., New York, 1955), p. 137.

[42] W. J. Schmidt, Z. wiss. Mikroskop. **52**, 158 (1935).

[43] F. S. Sjöstrand, J. Cellular Comp. Physiol. **33**, 383 (1949).

[44] F. S. Sjöstrand, J. Cellular Comp. Physiol. **42**, 15 (1953).

[45] J. B. Finean, F. S. Sjöstrand, and E. Steinmann, Exptl. Cell Research **5**, 557 (1958).

[46] F. S. Sjöstrand, Intern. Rev. Cytol. **5**, 455 (1956).

[47] H. Fernández-Morán, Exptl. Cell Research **5**, 586 (1958).

[48] T. H. Goldsmith and D. E. Philpott, J. Biophys. Biochem. Cytol. **3**, 429 (1957).

[49] J. J. Wolken, J. Capenos, and A. Turano, J. Biophys. Biochem. Cytol. **3**, 441 (1957).

[50] V. B. Wigglesworth, *The Principles of Insect Physiology*, (Methuen and Company, Ltd., London, 1950), fourth edition.

[51] W. H. Miller, J. Biophys. Biochem. Cytol. **5**, 421 (1957).

[52] K. von Frisch, *Bees: Their Vision, Chemical Senses and Language* (Cornell University Press, Ithaca, New York, 1950).

[53] H. Autrum and H. Stumpf, Z. Naturforsch. **5b**, 116 (1950).

[54] H. Autrum, Exptl. Cell Research **5**, 426 (1958).

[55] H. Fernández-Morán, Nature **177**, 742 (1956).

[56] B. Chance and G. R. Williams, Advances in Enzymol. **17**, 65 (1956).

[57] D. E. Green in *Harvey Lectures* (Academic Press, Inc., New York, 1958), Vol. LII, p. 177.

[58] A. L. Lehninger, Ch. L. Wadkins, C. Cooper, Th. M. Devlin, and J. L. Gamble, Science **128**, 450 (1958).

[59] A. Szent-Györgyi, *Bioenergetics* (Academic Press, Inc., New York, 1957).

[60] S. W. Peterson and H. A. Levy, Phys. Rev. **92**, 1082 (1953).

[61] B. Commoner, J. J. Heise, B. Lippincott, R. E. Norbert, J. Passoneau, and J. Townsend, Science **126**, 57 (1957).

[62] H. Fernández-Morán, Bol. acad. cienc. (Caracas) **51**, 1 (1953).

[63] G. Wald, Exptl. Cell Research, **5**, 389 (1958).

[64] T. H. Bullock, Federation Proc. **12**, 666 (1953).

[65] W. Shockley, *Electrons and Holes in Semiconductors*, Bell Laboratories Series (D. Van Nostrand Company, Inc., New York, 1953).

[66] D. D. Eley, G. D. Parfitt, M. J. Perry, and D. H. Taysum, Trans. Faraday Soc. **49**, 79 (1953).

[67] D. D. Eley and G. D. Parfitt, Trans. Faraday Soc. **51**, 1529 (1955).

[68] D. W. Pashley, J. W. Menter, and G. A. Bassett, Nature **179**, 752 (1957).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

39

Fine Structure of Lamellar Systems as Illustrated by Chloroplasts*

ALAN J. HODGE

Department of Biology, Massachusetts Institute of Technology, Cambridge 39, Massachusetts

THE fine structure of specialized organelles such as mitochondria, which are known to be concerned with the metabolic side of the carbon cycle, has been discussed in the three foregoing chapters (Bennett, p. 297; Sjöstrand, p. 301; Fernández-Morán, p. 319). The structures concerned with the other half of this cycle—namely, the photosynthetic aspect—are considered here.

It is of interest, in terms of the evolution of specialized photosynthetic organelles, to examine the fine structure of certain primitive types of cells, since these are presumably representative of organisms arising very early in the evolutionary process. Such a group of organisms are the blue-green algae. Examination in the electron microscope shows that, in certain types such as *Nostoc* (Fig. 1), there are no specialized membrane-bounded organelles such as mitochondria, nuclei, and chloroplasts; rather there exists a generalized membrane structure ramifying throughout the cytoplasm. It seems reasonable to conclude that, in such cells, the various metabolic functions such as oxidative phosphorylation and photosynthesis are carried on in relation to an apparently undifferentiated membrane system (or perhaps in specialized patches of the membrane), rather than in well-defined organelles. In other blue-greens (e.g., *Anabena*), the photosynthetic material is present in a more specialized and segregated state, but the primitive organelles lack limiting membranes. It is of interest to note that, in higher organisms, an intracellular membrane system (the endoplasmic reticulum)[1] appears to be universally present, and to have clear continuity with the membranous envelope of the nucleus (and possibly with the limiting membranes of other organelles). It seems likely, therefore, that comparative study of such primitive cells might yield important clues concerning the origin and interrelationship of the endoplasmic reticulum and specialized organelles such as mitochondria, chloroplasts, and nuclei in the cells of higher organisms.

Proceeding a little higher in the evolutionary scale to a typical green alga, such as *Nitella* (Fig. 2), one observes typical membranous elements of the endoplasmic reticulum, the characteristically structured mitochondria found in all higher organisms, and chloroplasts, bounded by a well-defined, double limiting membrane and containing a number of relatively well-ordered dense lamella within a finely granular matrix material.[2] Chloroplasts of this structural type are highly characteristic of the lower forms of plant cells.[3–6]

In the higher plants, of which corn (*Zea mays L.*) is chosen as an example, the structure of the chloroplasts[7–10] usually differs from the more primitive pattern already described. Figure 3 shows in transverse section the leaf of a three-week-old corn plant. Around each vascular bundle are a number of parenchyma sheath cells, the chloroplasts of which are specialized for the formation and storage of starch. Most of the photosynthetic activity, however, is carried out in the chloroplasts of the mesophyll cells. Even at the light-microscopical level, there are obvious differences in appearance between the chloroplasts of the parenchyma-sheath cells and those of the mesophyll cells. In the electron microscope (Fig. 4), the parenchyma-sheath chloroplasts of *Zea*[7] resemble algal chloroplasts (cf. *Nitella*, Fig. 2) in their over-all plan of organization. Within a double external limiting membrane, a number of densely staining lamellae (each about 130 A thick) are set in a finely granular matrix which presumably contains most of the soluble enzymes of the chloroplast.

The mesophyll chloroplasts are similarly lamellated but, in addition, possess well-defined regions (the grana) in which the lamellae are more densely and regularly packed (Fig. 4). If the section is in the plane of the lamellae, the grana appear as circular profiles. If, on the other hand, the plane of the section is normal to the lamellar plane, the grana appear as regular rectangular-shaped regions. The concentration of lamellar surface within the grana is about twice as high as in the intervening (intergrana) regions, and it can be seen (Fig. 5) that this comes about because of a pairing or bifurcation of the lamellae at the edges of the grana. This type of chloroplast with well-defined grana is characteristic of higher plants, most of which do not possess plastids of the type found in the parenchyma-sheath cells of *Zea*.

In both types of chloroplasts, the individual lamellae (each 130 A thick) of both the grana and intergrana regions exhibit a compound-layer structure (Figs. 6 and 7), consisting of a central dense line (the *P* zone, about 40 A thick and often resolvable as a doublet) on both sides of which are less dense layers (the *L* zones). Finally, the entire compound layer structure is edged by very thin dense lines, the *C* zones. Such differences in density arise in part from differences in reactivity

* Most of the work described herein was carried out in the Chemical Physics Section, C. S. I. R. O., Melbourne, Australia, in collaboration with Dr. F. V. Mercer and Dr. J. D. McLean of the Botany Department, University of Sydney, N. S. W.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

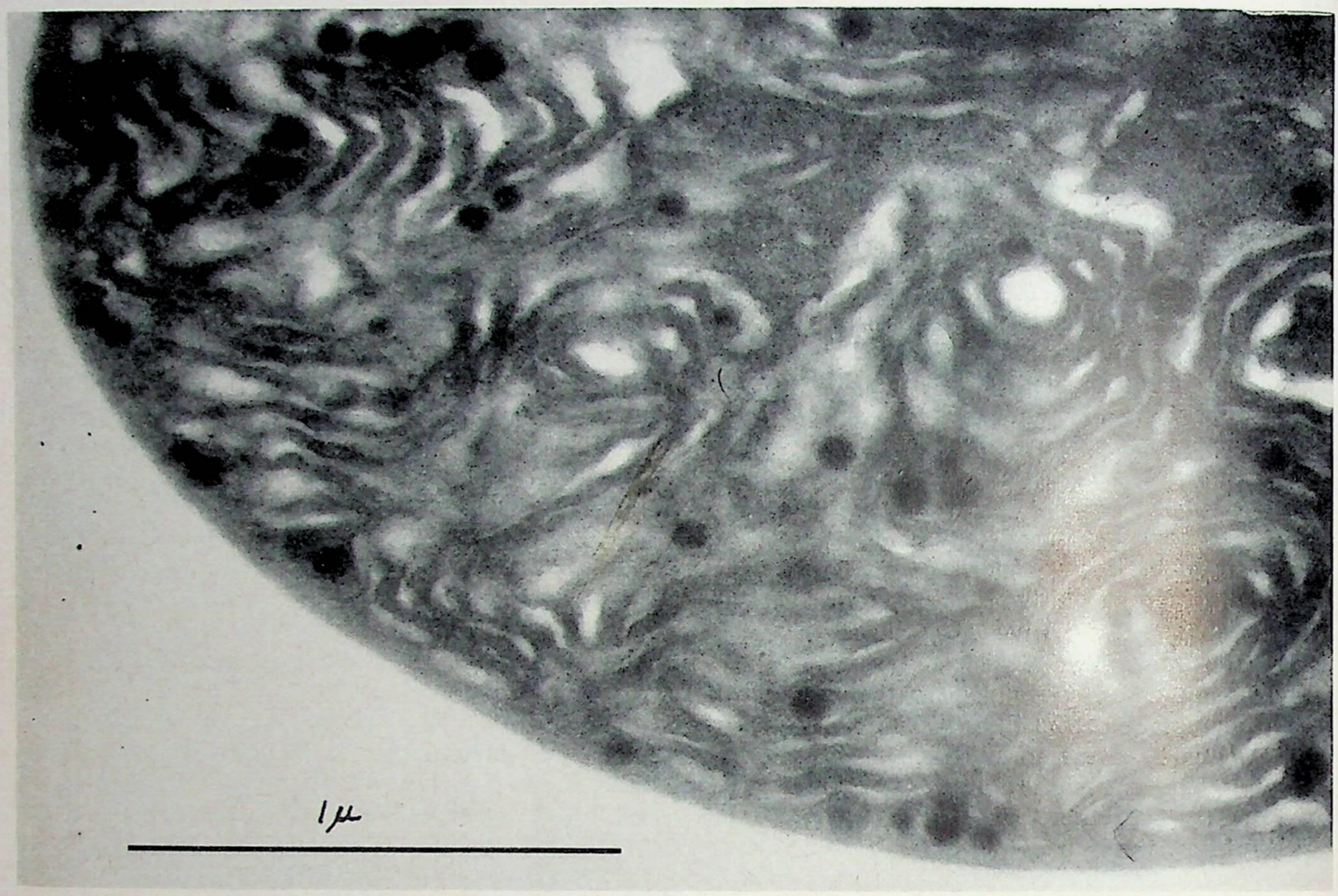


FIG. 1. Vegetative cell of a *Nostoc* spp., showing the generalized whorl-like lamellar system and the absence of specialized organelles such as chloroplasts and mitochondria. ×60 000. (Unless otherwise indicated, all illustrations are electron micrographs of thin sections of plant material fixed in osmium-tetroxide solutions of appropriate *p*H and tonicity, and embedded in methacrylate. Details are given in references 2, 7, 13, and 14.)

with osmium tetroxide as well as from intrinsic differences in electron density within the various components of the structure. Within the grana, the close packing of such compound lamellae results in close apposition of *C* zones and, thus, gives rise to a layer structure (Figs. 8 and 9) with a repeat period of 130 A, bearing a remarkable resemblance to that found in the myelin sheath of nerve fibers. The central dense lines of the compound lamellae correspond to the major dense lines of the periodic structure within the grana and are often observed as doublets (Fig. 9), each line of the pair being about 15 A thick. It is clear, therefore, that the symmetrical structure of the individual compound lamellae arises from the close apposition of two structurally asymmetric "unit membranes" (each about 70 A thick) as shown schematically in Fig. 11. The *I* zones, which occur midway between *P* zones in the grana, arise by apposition of the *C* zones of the unit membranes. The work of a number of people, notably Robertson,[11] indicates that the cell membranes and intracellular membrane systems of most, if not all, cells consist of such unit membranes in single, double, or compound array. This also appears to be true of plant cells. In Fig. 10, the tonoplast is seen as a single membrane with a pair of dense edges, and the double limiting membrane of the chloroplast consists of two such unit membranes. The intrinsic asymmetry of these unit membranes is not usually evident in the electron microscope except where they are stacked in double or multiple array, as in the lamellae and grana of chloroplasts and in myelin sheath. In both cases, this stacking results in a set of major dense lines with less-dense intermediate lines in between. As already indicated by Fernández-Morán (p. 319), there is good reason to believe that each 70-A membrane comprises a double layer of mixed lipids (the *L* zones) sandwiched between two thin monolayers of protein which stain densely with osmium tetroxide, with a probable contribution to this latter density arising from reaction of the hydrophilic groups of the phospholipids with osmium tetroxide. In the case of myelin sheath, a very satisfactory correlation between electron-microscopic and x-ray diffraction data has been achieved[12] and the chemical composition is reasonably well known. It seems fairly certain, therefore, that this type of structure is essentially correct for most cellular membranes, with only such minor differences in detail arising as are demanded by differences in function.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

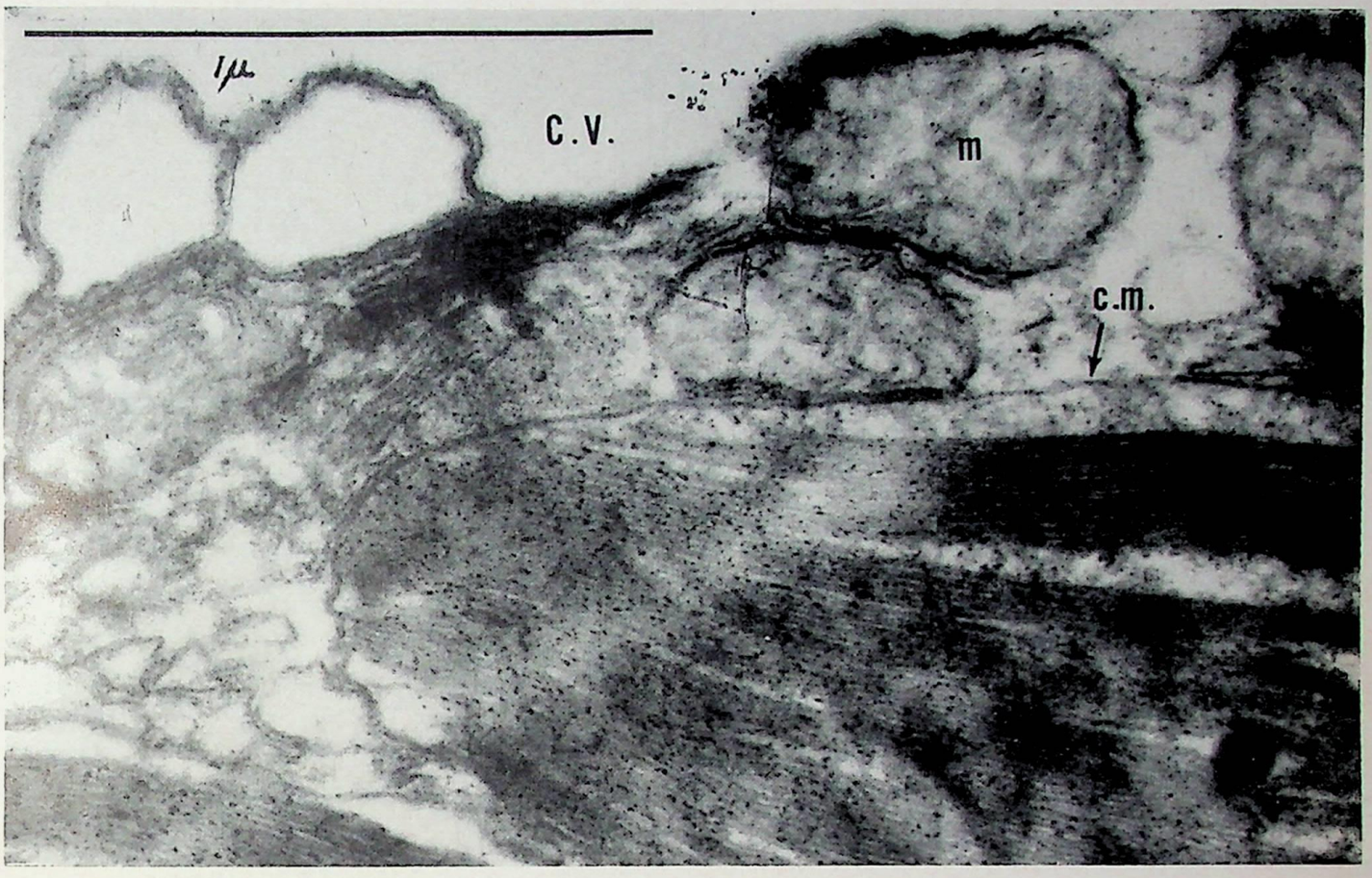


FIG. 2. Transverse section through the cytoplasm of a *Nitella* cell, showing the lamellated structure of a chloroplast surrounded by its limiting membrane *c. m.*, mitochondria *m*, various membranous elements within the cytoplasm, and the tonoplast lining the lumen of the central vacuole *C. V.* ×75,000.

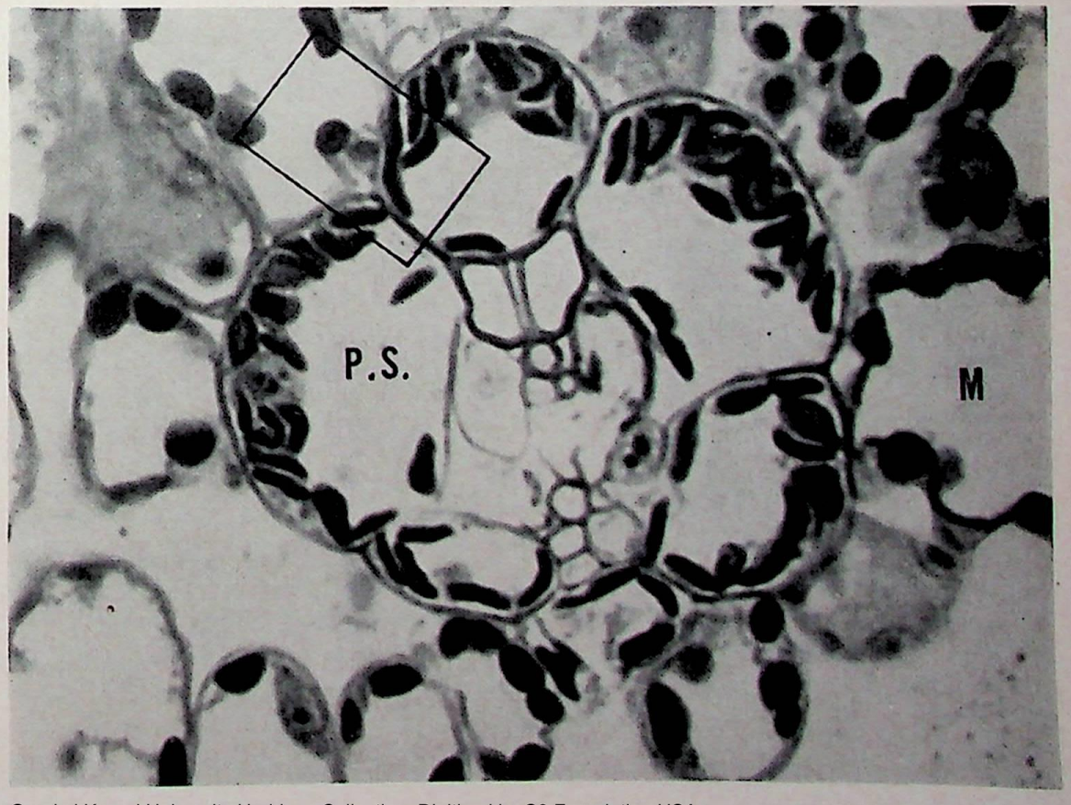


FIG. 3. Phase-contrast micrograph of a relatively thick transverse section through the leaf of a three-week-old plant of *Zea mays L.*, illustrating the difference in appearance between the chloroplasts of the parenchyma-sheath cells *P. S.* and those of the mesophyll cells *M*. A vascular bundle lies in the center of the field.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

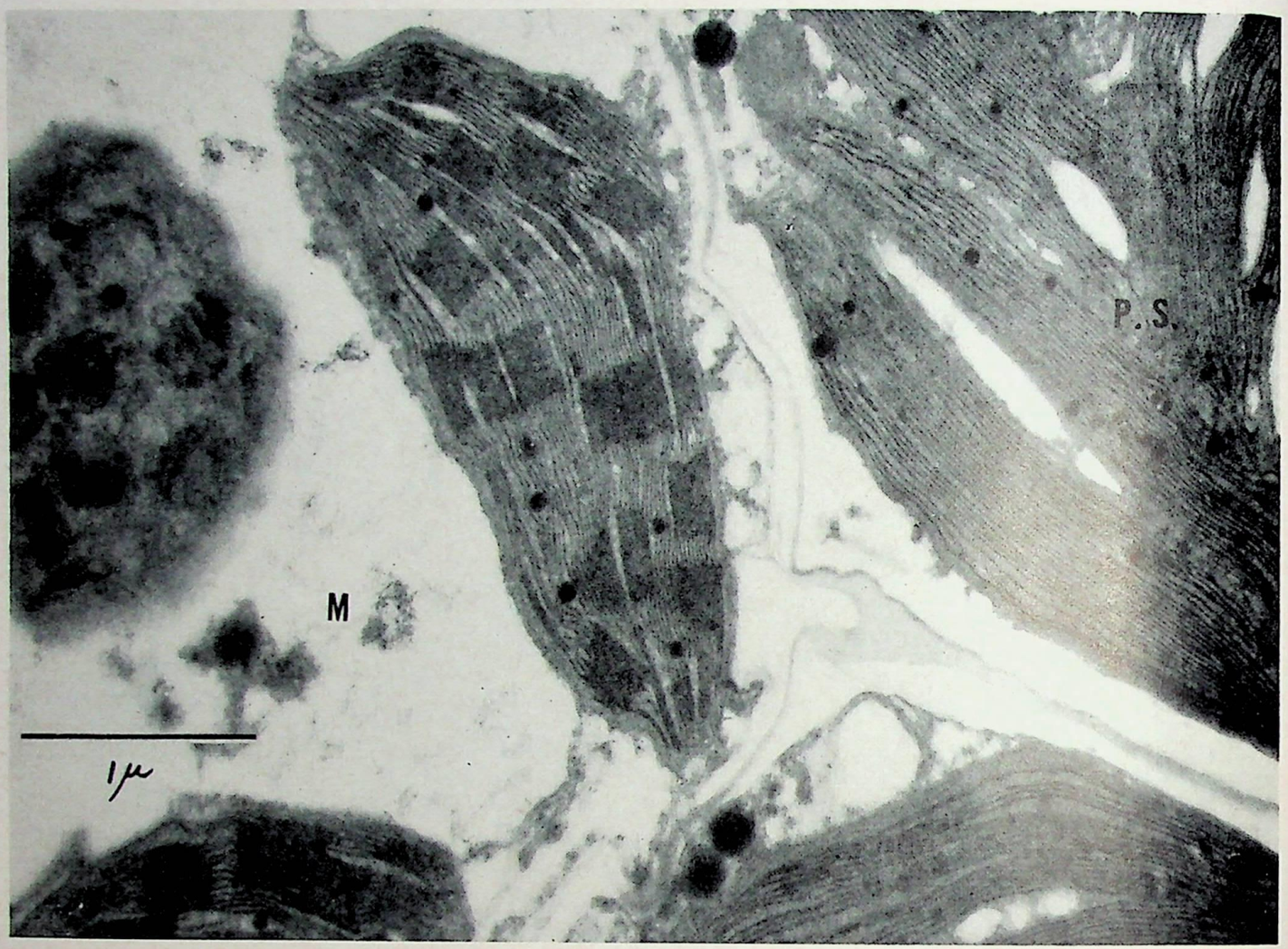


FIG. 4. Electron micrograph of a region similar to that outlined in Fig. 3, illustrating the lamellar structure both of parenchyma-sheath cells *P. S.* and mesophyll *M* chloroplasts of *Zea*. In the mesophyll chloroplasts, the grana appear as dense rectangular regions when the section is normal to the lamellar plane, and as circular regions when the plane of section is parallel to the lamellar plane. ×30 000.

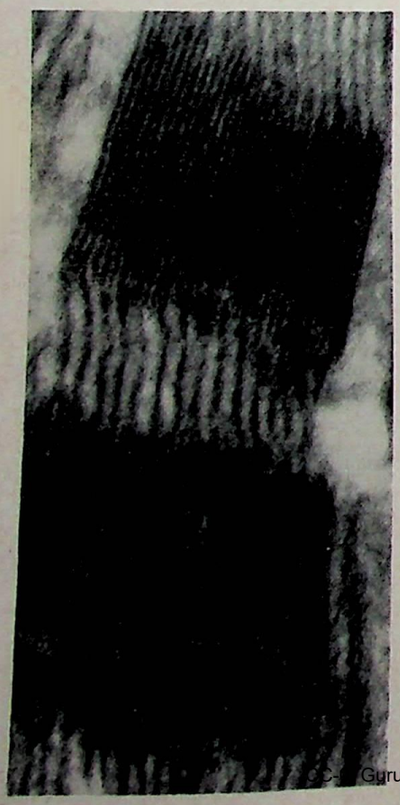


FIG. 5. Two grana in a mesophyll chloroplast of *Zea*, showing the type of connection between the grana lamellae and those in the intervening (intergrana) regions. (Cf. Fig. 11.) ×85 000.

In the case of the chloroplast, the situation is less well documented, but certain conclusions can be drawn from the available structural and chemical data, and by extrapolation from the better characterized myelin system. It can be seen from Fig. 10 that the membranes making up the lamellae and grana of the chloroplast are considerably more dense than the external limiting membranes of the chloroplast and the tonoplast, which exhibit densities more characteristic of the usual cellular membrane systems. In particular, the *L*-zones of the chloroplast lamellae exhibit higher densities than are found in cell membranes, endoplasmic reticulum, etc. A possible explanation of this may lie in differences of chemical composition. It is known that the chloroplast is deficient in phospholipids as compared with myelin sheath.[7] This deficiency is counterbalanced by the presence of considerable amounts of carotenoids and chlorophyll. Furthermore, as is shown later, there is good evidence to suggest that chlorophyll is an integral component of the chloroplast lamellae and that sufficient is present for its incorporation as a monolayer over the entire lamellar area of the chloroplast as estimated from electron micrographs. Thus, the general conclusion

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

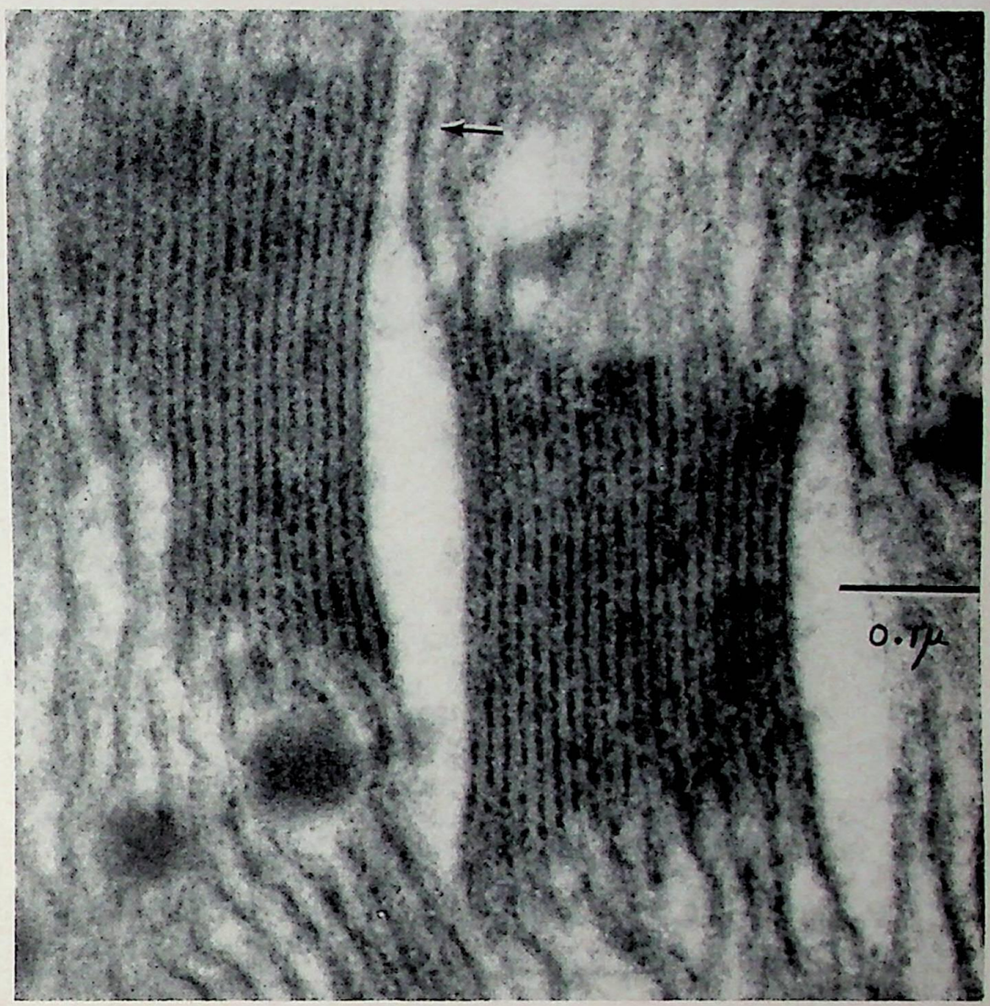


FIG. 6. Mesophyll chloroplast of *Zea*, illustrating the high degree of order within the grana, and the compound-layer structure of the intergrana lamellae (arrow). ×160 000.

seems justified that the photosynthetic lamellar elements of the chloroplast consist of structurally asymmetric unit membranes (Fig. 11) in which chlorophyll and carotenoids are incorporated in orderly array. It is of interest to note that the exciton-migration theory of energy transfer proposed by Calvin appears to require the presence of such a structural asymmetry. The type of model suggested by Calvin (p. 157) has the necessary asymmetry (Fig. 12) and appears to be consistent with the presently available chemical and structural evidence. Such a staggered, partially overlapping configuration of the chlorophyll molecules has the further advantage over other models (which usually involve coplanarity of the porphyrin "heads" of the chlorophyll) in that it offers a plausible explanation for the low values of dichroic effects so far observed in chloroplasts. Furthermore, it would allow a greater degree of π-electron interaction than allowed by the coplanar type of model, thus facilitating energy transfer, and is consistent with the type of packing found in crystals of polycyclic organic compounds by x-ray diffraction. As has been seen the chloroplasts of higher plants characteristically contain grana in which the unit membranes are arranged in a highly ordered and close-packed array. The significance of this specialization is unknown at the present time, but it is tempting to speculate that the high degree of order might lead to more efficient trap-

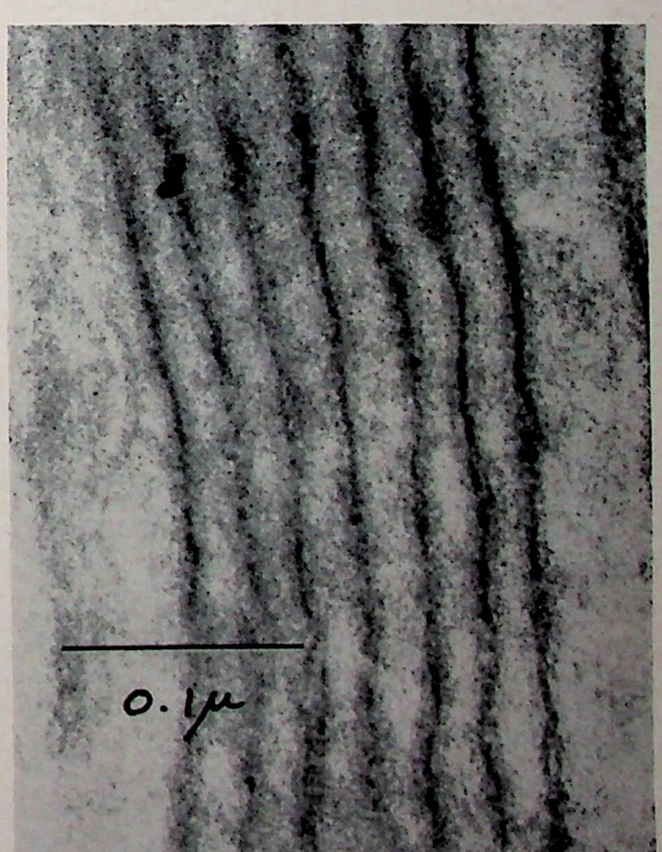


FIG. 7. Compound layer structure of the lamellae in a parenchyma-sheath chloroplast of *Zea*. ×270 000.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

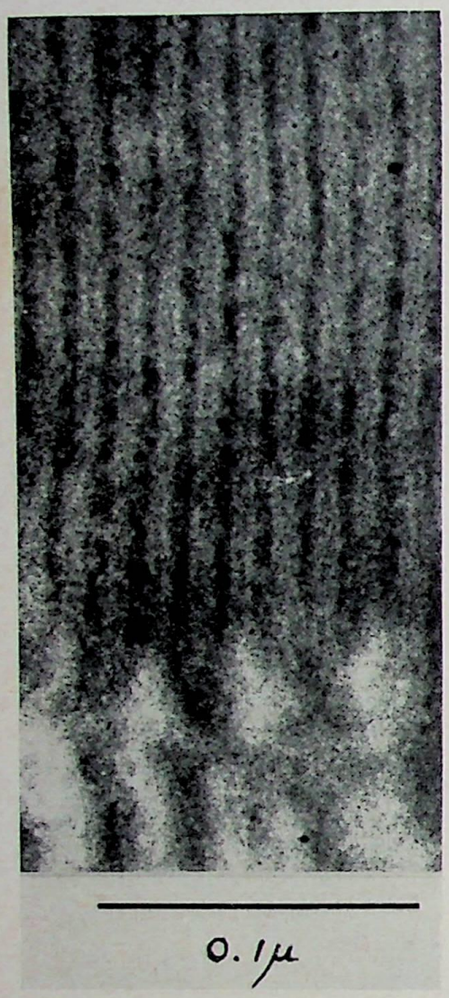


FIG. 8. Mesophyll chloroplast of *Zea* to illustrate the myelin-like structure within a granum. Note the main periodicity of about 130 A and the presence of fainter intermediate lines. ×370 000.

ping of photons, or perhaps result in facilitation of exciton migration as a result of a stabilization of the layer structures which possess the conduction bands necessary for energy transfer.

SWELLING PROPERTIES OF THE CHLOROPLAST

The swelling characteristics of chloroplasts offer a striking demonstration of the "plasticity" of such lipoprotein layer systems. As already mentioned the swelling of mitochondria can be largely prevented by active expenditure of energy from ATP hydrolysis. A similar phenomenon exists in the case of chloroplasts. Klein (unpublished work) has shown that the swelling response of isolated chloroplasts depends on whether the suspension of plastids is illuminated or is kept in the dark.

When isolated *Nitella* chloroplasts are placed in a hypotonic medium, swelling results first in a separation of the lamellae to distances many times that characteristic of the intact chloroplast.[2] In highly hypotonic media, the lamellae break up. The hydrophobic hydrocarbon chains of the lipids thus are exposed at the free edges of the membranes, giving rise to an unstable situation, with the result that such free edges "zip" together to form membrane-bounded vesicles. Similar results have been obtained for isolated *Zea* chloroplasts (Fig. 13), and the characteristically vesicular structures found in so-called microsome fractions isolated by conventional procedures arise in similar fashion (e.g., Hodge[13]) from the endoplasmic reticulum. This plastic behavior is understandable in terms of what is known of membrane structure. In the type of layer structure under consideration here, it seems likely that the oriented lipid molecules possess a high degree of rotational freedom and considerable translational freedom within the plane of the membrane. Both properties are characteristic of the smectic fluid-crystal state and confer a remarkable capacity for changes in the topographical configuration of such membrane systems.

CHLOROPLAST DEVELOPMENT

The development of the chloroplast is of interest in that it appears to take place by a process which is essentially the reverse of that involved in the swelling phenomena just discussed. In brief, the compound- and extended-layer structures of the lamellae and grana appear to be formed by a process involving the fusion of small vesicles or micelles one with another to form extended cisternae or "double-membrane structures."

The development of a leaf or of a plant is a complicated sequence of events, the time and spatial relationships of which lead to difficulty in following the process of chloroplast development. The etiolated plant (one grown from seed in total darkness) is a much more

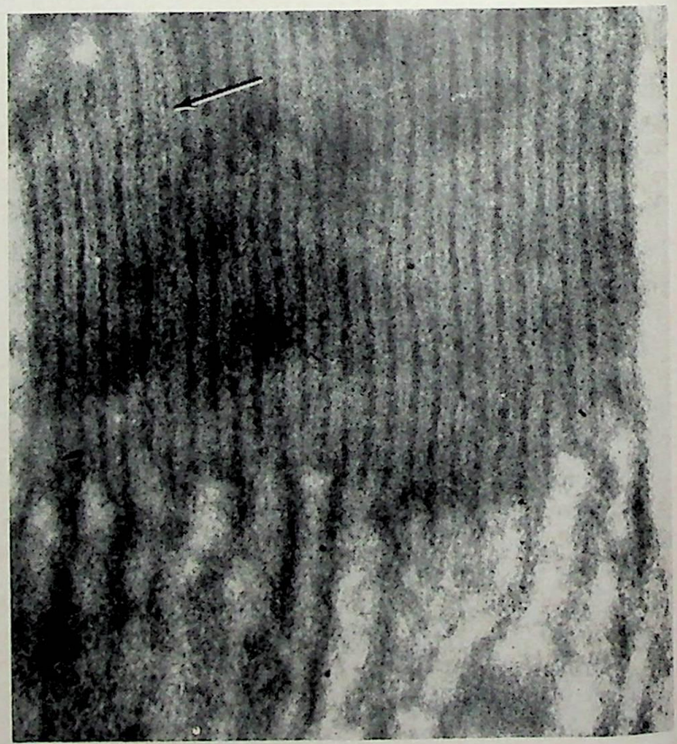


FIG. 9. A granum in a mesophyll chloroplast of *Zea*, illustrating the splitting of the main dense lines (arrow). The central dense lines (*P* zones) of the intergrana lamellae are also occasionally resolvable as doublets, thus showing that the compound lamellae consist of two closely apposed asymmetric unit membranes. ×220 000.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

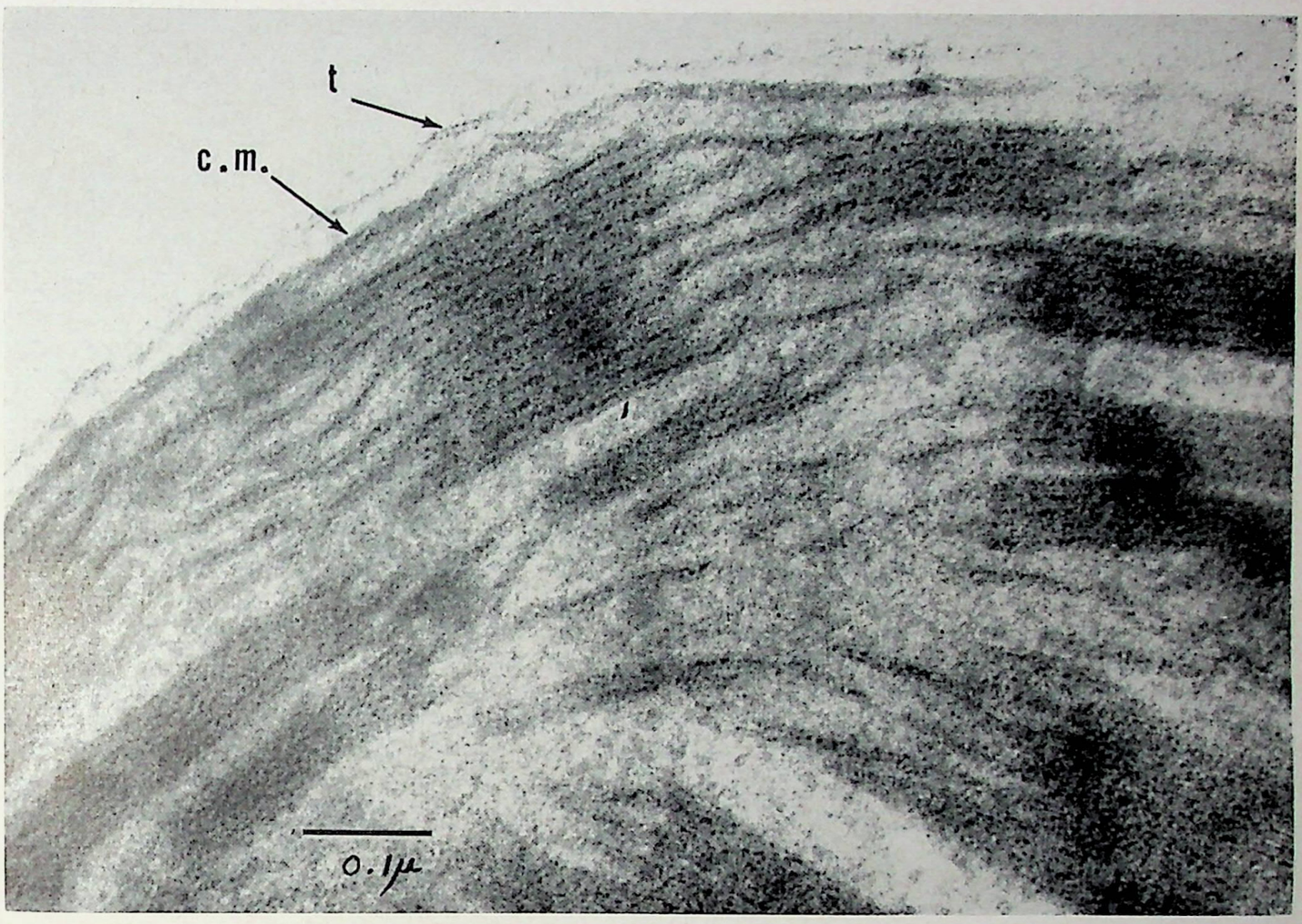


FIG. 10. Chloroplast in a three-week-old wheat plant, showing lamellar structure very similar to that in *Zea mays*. The tonoplast *t* appears as a single membrane, and where suitably oriented with respect to the plane of the section appears as two fine dense lines with a less dense layer between them, the over-all thickness being about 70 A. The chloroplast envelope *c. m.* comprises two unit membranes spaced about 100 A apart. ×160 000.

favorable system for such a study, since such seedlings can be exposed to light for various periods and the effects on chloroplast structure noted at definite time intervals following such exposure. This system depends on the fact that light is essential for chlorophyll synthesis in higher plants.

Figure 14 shows typical plastids in an etiolated leaf of *Zea*.[14] In the absence of chlorophyll, the plastids, although recognizable as well-defined organelles with external limiting membranes, fail to develop lamellae. Instead, the interior of each plastid is partially filled with a mass of small vesicles (the prolamellar body).

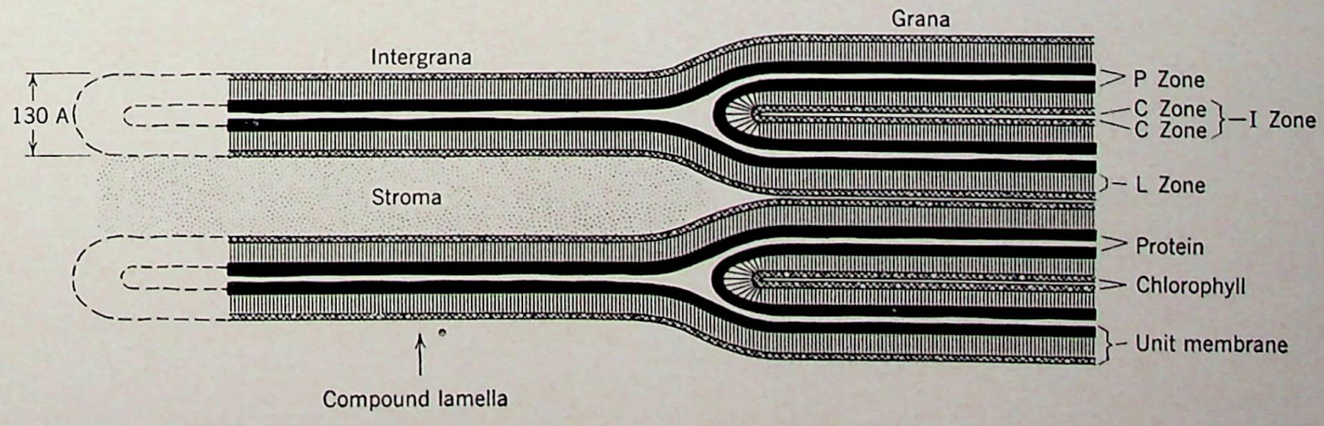


FIG. 11. Diagram to illustrate the densities observed in electron micrographs of osmium-fixed chloroplast lamellae and grana, and the structureal relationships involved in the formation of the compound lamellae (and grana) from structurally asymmetric unit membranes.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

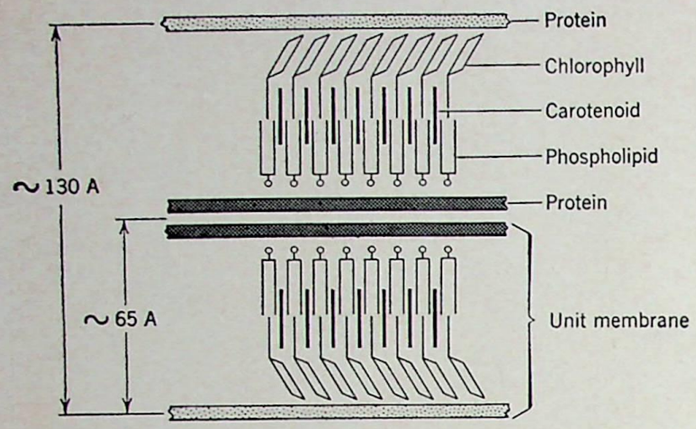


FIG. 12. Diagram to illustrate one way in which the compound lamellae observed in electron micrographs of chloroplasts could be built up from structurally asymmetric unit membranes having a molecular structure of the type proposed by Calvin (p. 157).

On placing the plant in daylight, chlorophyll is synthesized (as evidenced by a progressive greening of the leaves), and definite structural changes are observed in the plastids. Double membranes are observed (Fig. 15) apparently emerging from the prolamellar body.[14] At a later stage, typical chloroplast lamellae are seen, usually in a pattern radiating from a progressively decreasing prolamellar body (Fig. 16). It is of interest to note that in certain *Zea* mutants which are unable to synthesize chlorophyll, the plastids are similar to those of etiolated normal *Zea* in that they lack formed lamellar elements and contain only small vesicular structures. The evidence thus strongly suggests that chlorophyll is at least essential for the formation of typical chloroplast lamellae and is probably an integral component of these membrane structures.

At a later stage, rudimentary grana with a fine structure indistinguishable from that of the more mature grana already described may be observed (Fig. 17) within the mesophyll chloroplasts. After two days, the developing plastids closely resemble normal chloroplasts and no prolamellar bodies can be discerned, presumably as a result of complete conversion of the small vesicles into formed lamellar elements. At this stage and in earlier stages of recovery from etiolation, however, and in young chloroplasts of normal plants, one commonly observes immediately under the chloroplast limiting membrane a number of vesicles and sacs (Fig. 18) with membrane densities considerably lower than those characteristic of the already differentiated lamellae and grana. These observations suggest the possibility that this region, or the limiting membrane of the chloroplast itself, may be responsible for the formation of the small vesicles, perhaps by outpocketing and pinching-off from the limiting membrane, or by some direct synthetic process.

The concept of vesicle fusion either with other vesicles or with already existing extended membranes is attractive in relation to (1) the formation and maintenance of structures of higher order, such as the

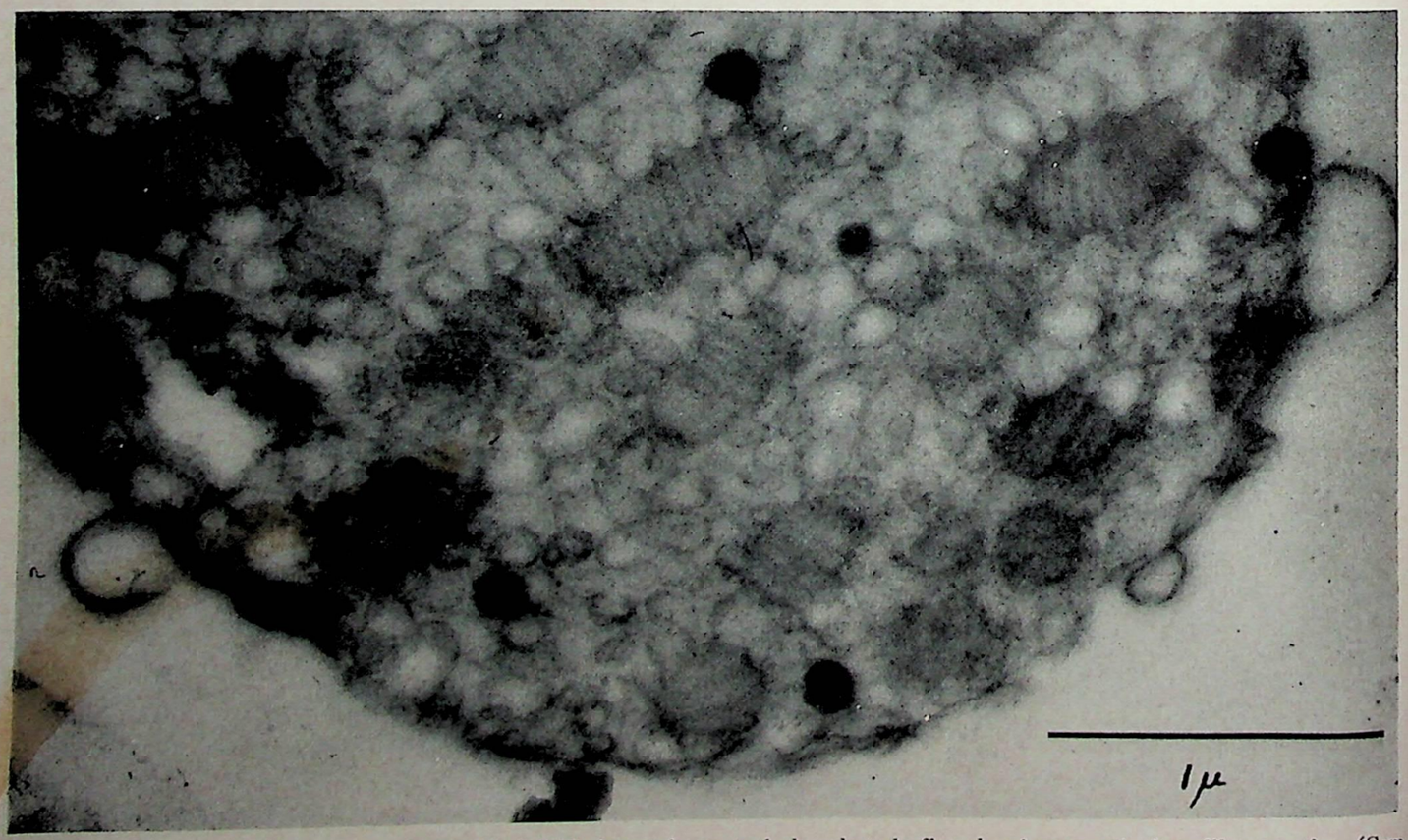


FIG. 13. Mesophyll chloroplast of *Zea* isolated in 0.5 M glucose in neutral phosphate buffer, showing a marked swelling reaction. (Compare with Figs. 4 and 6.) Note that the intergrana lamellae have broken up and formed large numbers of vesicles. × 40 000.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

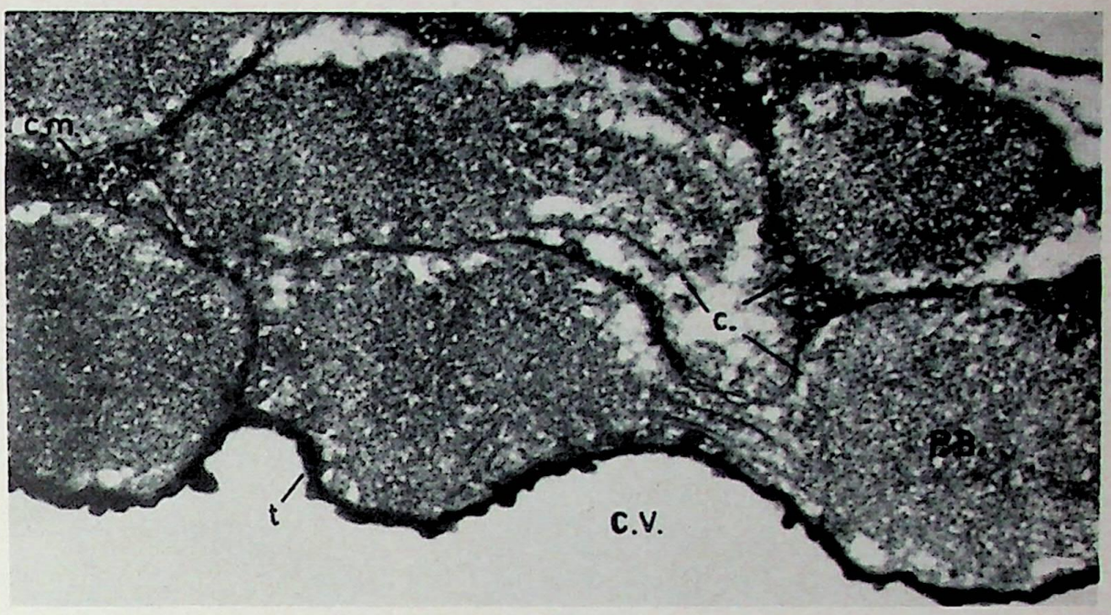


FIG. 14. Plastids in an etiolated leaf of *Zea*. The plastids contain masses of minute vesicles, but are devoid of lamellae, presumably because the absence of chlorophyll prevents their formation.

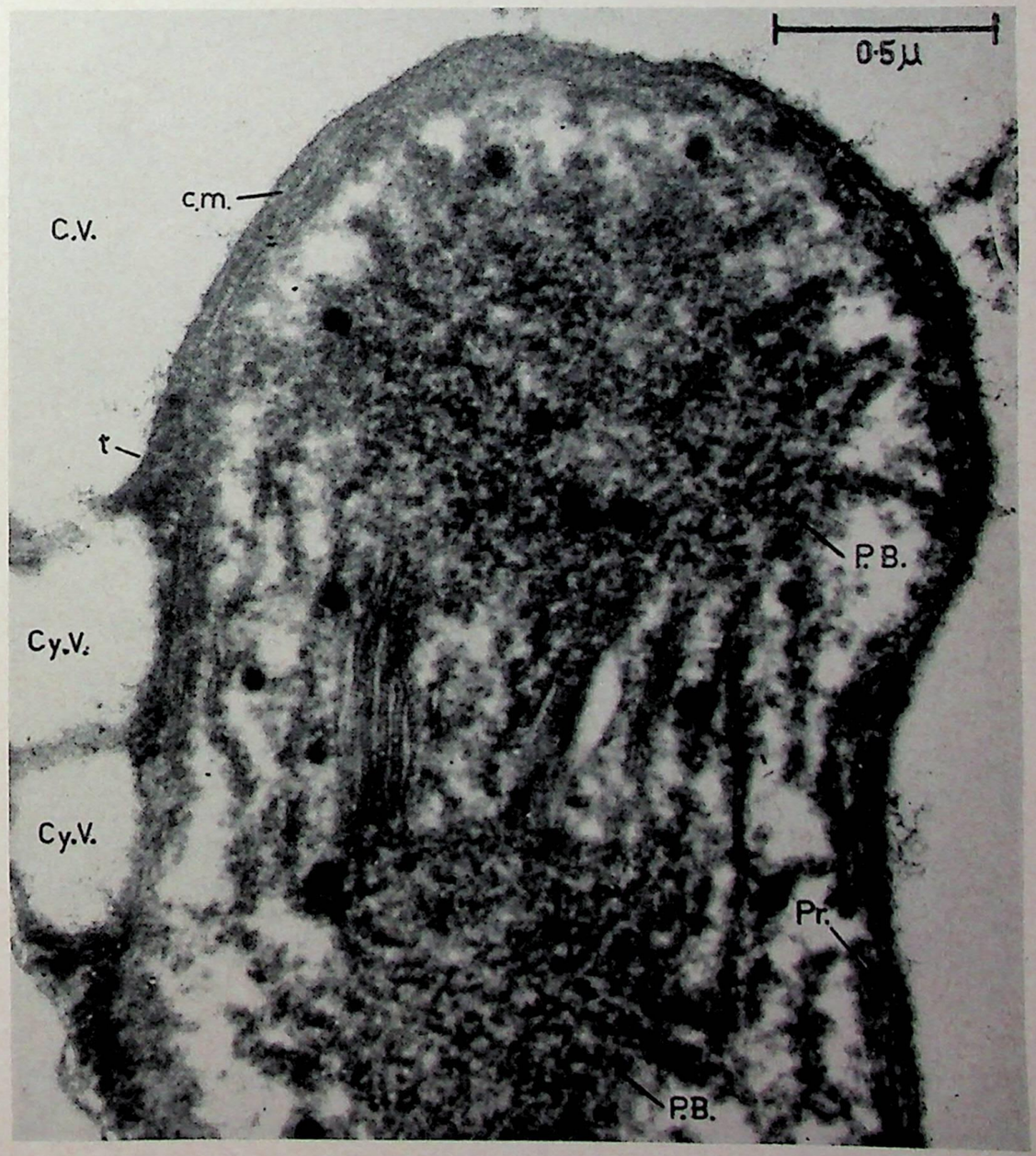


FIG. 15. Plastid from an etiolated leaf of *Zea* after several hours exposure to daylight showing an early stage in the formation of lamellae. Note that the lamellae appear to arise initially as double membranes, apparently by a process involving an orderly fusion of the small vesicles comprising the prolamellar body *P. B.* ×55 000.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

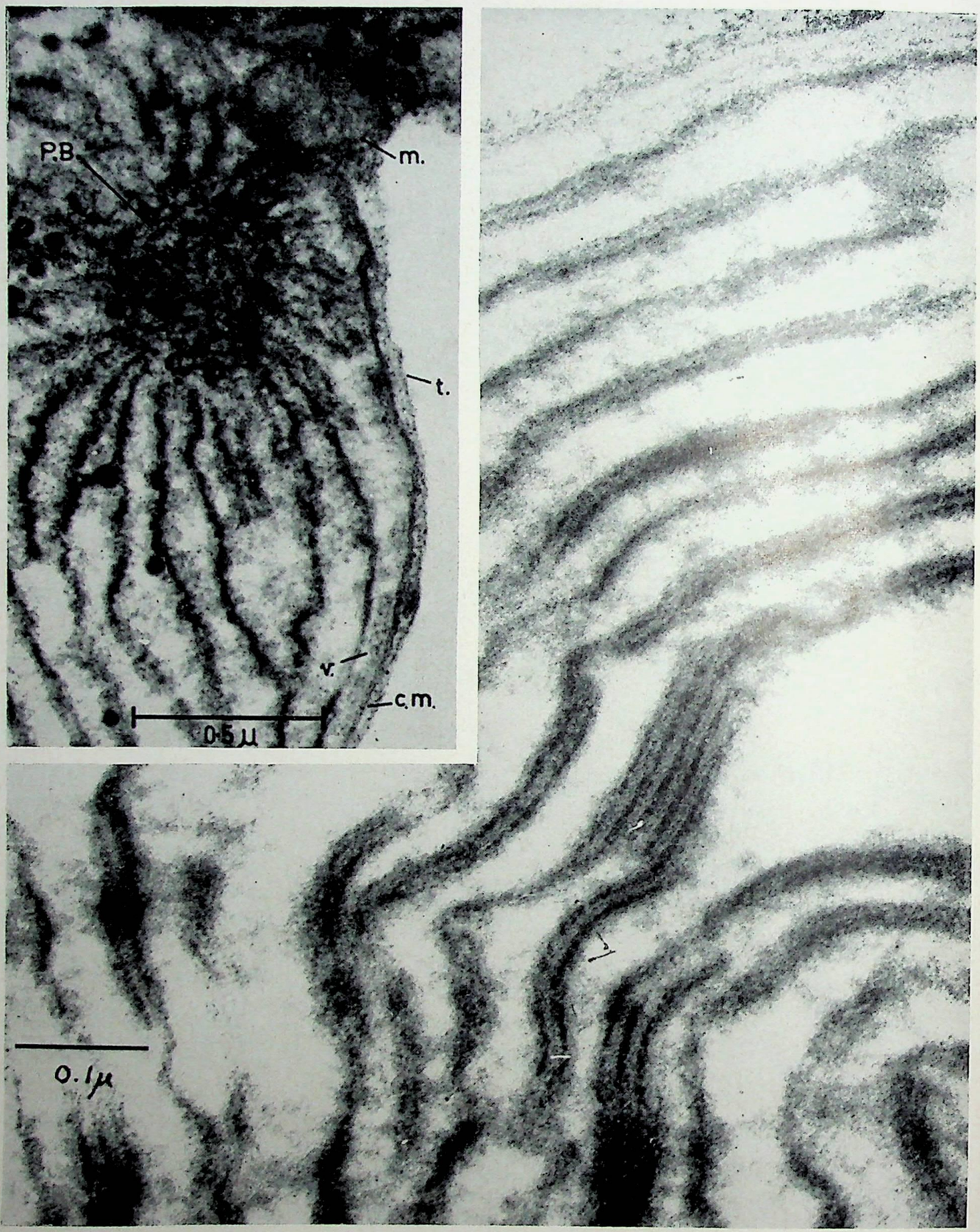


FIG. 16 (upper left). A somewhat later stage in recovery from etiolation than that shown in Fig. 15, illustrating the progressive reduction in size of the prolamellar body *P. B.* and formation of typically dense chloroplast lamellae.

FIG. 17 (upper right and lower area). Mesophyll plastid of etiolated *Zea* after 20-hr exposure to daylight showing the beginning of grana formation. ×210 000.

endoplasmic reticulum and the various organelles of the cell (the vast increase in membrane area required for the formation of the myelin sheath of nerve fibers could be accounted for by fusion of vesicles with the Schwann-cell membrane); and (2) the transport both into and out of the cell of various substances. Such a

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

FIG. 18. Mesophyll plastid of etiolated *Zea* after 48-hr exposure to daylight. Typically dense chloroplast lamellae and grana are present. Note the presence of less-dense vesicles and cisternae in the region immediately beneath the chloroplast envelope *c. m.* ×85 000.

mechanism has been proposed for the extrusion of acetylcholine in "quantized" amounts from the nerve endings in myoneural junctions, and for the passage of zymogen granules from the apical regions of acinar pancreas cells through the limiting membrane into the lumen of the acinus. Similarly, the occurrence of the reverse process (i.e., pinocytosis) seems to be well documented. While mechanisms of this general type are widely supported by indirect experimental evidence, and appear to be thermodynamically feasible in terms of current knowledge concerning the structure of lipoprotein layer systems, the crucial problem of energy coupling in the physical control of such membrane systems remains unelucidated.

BIBLIOGRAPHY

[1] K. R. Porter in *Harvey Lectures* (Academic Press, Inc., New York, 1957), Vol. LI.

[2] F. V. Mercer, A. J. Hodge, A. B. Hope, and J. D. McLean, Australian J. Biol. Sci. **8**, 1 (1955).

[3] R. Sager and G. E. Palade, Exptl. Cell Research **7**, 584 (1954).

[4] J. J. Wolken and G. E. Palade, Ann. N. Y. Acad. Sci. **56**, 873 (1953).

[5] P. A. Albertsson and H. Leyon, Exptl. Cell Research **7**, 288 (1954).

[6] H. Leyon, Exptl. Cell Research **6**, 497 (1954).

[7] A. J. Hodge, J. D. McLean, and F. V. Mercer, J. Biophys. Biochem. Cytol. **1**, 605 (1955).

[8] H. Leyon, Exptl. Cell Research **7**, 265 (1954).

[9] E. Steinmann and F. S. Sjöstrand, Exptl. Cell Research **8**, 15 (1955).

[10] M. Cohen and E. Bowler, Protoplasm **42**, 4, 414 (1953).

[11] J. D. Robertson, J. Biophys. Biochem. Cytol. **4**, 39 (1958).

[12] H. Fernández-Morán and J. B. Finean, J. Biophys. Biochem. Cytol. **3**, 725 (1957).

[13] A. J. Hodge, J. Biophys. Biochem. Cytol. Suppl. **2**, 221 (1956).

[14] A. J. Hodge, J. D. McLean, and F. V. Mercer, J. Biophys. Biochem. Cytol. **2**, 597 (1956).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

40
Membrane Transport

ROBERT W. BERLINER

National Heart Institute, National Institutes of Health, Public Health Service, U. S. Department of Health, Education, and Welfare, Bethesda 14, Maryland

IN discussing biological membranes one deals with a considerably higher level of organization than has characterized most of the material covered in the foregoing papers, even in the consideration of cellular biology, since one must deal with everything from the presumed limiting membranes of individual cells to entire layers of cells, each of considerable complexity and, in some cases, even consisting of several different cell types. However, in general, the problem is oversimplified here by disregarding whenever possible all details of structure which are not obviously concerned directly with the processes under immediate examination.

The animal body can be considered to be made up of a number of compartments separated one from another and from the external environment by a series of membranes. Each compartment has a more or less characteristic composition, often differing markedly from that of the adjacent compartment so that across these membranes there may be rather steep gradients of chemical concentration and often of electrical potential as well. The problem is to clarify the means by which various materials cross these membranes and by which the striking gradients are maintained. To some extent the penetration of various materials can be defined in terms of relatively simple forces; in other instances it is apparent that the movements of certain substances require the utilization of energy, a process which has been designated as active transport and which is currently the subject of numerous studies. In no case, however, can it be said that the mechanisms involved have been identified.

Consider, in a general fashion, the possible ways in which materials might cross biological membranes and the forces which might effect these movements. These considerations are based heavily upon the treatment of the subject of membrane transport as developed by Ussing. In general, the definable forces which produce movement across membranes can be classified as: (1) those due to gradients of chemical activity including differences in concentration or activity coefficients; (2) those due to gradients of electrical potential; and (3) those exerted upon solutes by the flow of solvent through solvent-filled channels. In addition, one finds in practically all biological systems movements of solute which are explained by none of these definable forces and which, therefore, are designated as "active transport." The definition of active transport is a rather arbitrary affair. As defined, it has the advantage of singling out those solutes which must be considered directly involved in the process by which metabolic energy is utilized to perform transport work. However, clearly all concentration or electrical gradients and flows in response to pressure or osmotic gradients ultimately depend upon metabolic work even though this may be remote from the particular membrane under consideration. So the requirement of metabolic work is not sufficiently exclusive for an adequate definition of active transport.

However, the definition does not include all processes which have specificities beyond simple diffusion and flow and does not include the process by which certain substances can cross membranes only, say, by combining with a specific carrier even though the movement is entirely downhill with respect to gradients of electrochemical potential. One must recognize, therefore, the arbitrariness of this definition. Presumably, when, and if, a better understanding of the discrete processes involved is attained, the need for the term active transport will disappear.

Of course, the manner in which solutes may cross membranes depends upon the nature of the membrane. If it is a continuous lipid layer, substances can pass this layer only by dissolving in it, although their passage through such a layer might be facilitated by combination with some component of the lipid layer which would enhance the lipid solubility of the solute. For such membranes, the permeability to particular solutes should, in general, be a function of lipid solubility. Indeed, observations going back many years have shown that in several chemical series the permeation of cells can be related to lipid solubility.[1] However, there are many features of cellular membranes which cannot be explained if they are assumed to consist of a solid lipid layer and it is generally believed that the lipid layer contains pores, through which water forms a continuous phase and through which the water soluble solutes pass. Indeed, it is only through such channels that forces produced by the flow of solvent can act. Since, as is pointed out later, there is evidence that solvent flow does produce movement of solute, it is apparent that such solvent-filled pores (and in this case, of course, the solvent referred to is water) are present in some membranes. For present purposes, it is convenient to accept the view that cell membranes generally consist of lipid layers penetrated by aqueous pores, the walls of which may bear an excess of fixed charged groups which will impede the passage of ionic species of the same sign. In some cases where membranes consist of more complicated structures than the

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

plasma membrane of single cells and may consist of an entire layer of cells attached to a common basement membrane, there may be reason to believe that the passage of water soluble substances occurs between the cells rather than through them. This appears to be the case in blood capillaries,[2] but probably not in others such as the mucosal lining of the alimentary tract and the skin of frogs and toads.

Perhaps the best-studied biological membrane system and the one in which theoretical considerations have been applied most explicitly is the frog skin which, in particular, Ussing has used as a model system for extensive studies.[3–5] Unfortunately, from the structural point of view, the frog skin is rather complicated, but perhaps it may be simplified for present purposes by making the probably valid assumption that only the basal layer of cells is involved in the transport processes of concern here. The skin can be stripped rather easily from the abdomen of the frog and stretched *in vitro* as a diaphragm between two chambers. It will then survive in this isolated state, perform its metabolic operations, and retain a well-maintained capacity to perform active transport for some hours. The sheet of tissue thus obtained consists of a loose layer of connective tissue, a continuous basement membrane, and several layers of epithelial cells. As one proceeds from the basement membrane toward the outer skin surface, the cells progressively become more flattened and keratinized and, presumably, lose their metabolic activities. In any case, it may be assumed for present purposes that all of the observed phenomena are attributable to the basal layer of cells—an assumption to some extent supported by the fact that toad bladder has been shown to exhibit many similar properties but to have only the single layer of cells corresponding to the basal layer of cells in the skin.[6]

Now, it has been known for many years that the frog skin generates an electrical potential, the solution in contact with the inner surface being positive with respect to the outside. It has also been known that the frog in the intact state is able to extract salt from its surroundings even though the outside concentration of salt is far lower than in the tissue fluids of the animal. For instance, Krogh[7] found that the frog which has a concentration of sodium in its extracellular fluid of something over 100 mEq/L (milliequivalents per liter) was able to take up salt from its surroundings when the outside concentration was as low as one hundredth of a milliequivalent per liter—a concentration ratio of 10000:1. Furthermore, the uptake was completely specific for sodium—no potassium or calcium was taken up even when the concentration was many times higher. One other unusual observation was made by Hevesy *et al.*[8] when the technique of using isotopic tracers was introduced. They found that the frog sitting in water had a net uptake of water several times greater than predicted from the measured influx of tagged water and the assumption that the flux in each direction across the skin should be proportional to the water activity in the solution of origin—in this case, the lymph or extracellular fluid of the frog and the water in which the frog was immersed. Observations such as this one led to the hypothesis that water was actively transported across such membranes (similar observations were made by Visscher and his associates using the intestine of the dog[9])—however, a simpler physical explanation is shown to make the assumption of active water transport unnecessary. One further observation on the intact frog might be mentioned before going on to the more detailed observations—this is the fact that the rate of uptake of water from the surroundings can be enhanced greatly by injection of the frog with posterior pituitary extracts.[10] Hence, the term amphibian water-balance principle was introduced to describe this activity, which, however, has been since shown to be at least qualitatively reproduced by the use of the purified polypeptide hormones of the posterior pituitary.

These, than, are the gross observations which Ussing and his collaborators set about to dissect by the application of isotopic tracers. Disregarding for the time being the net movement of water which in any case is negligible when the skin is bathed on both sides with solutions of equal osmotic pressure, Ussing showed that the forces acting on an ion to cause movement across the intervening membrane would produce a flux from one surface to the other described by the equation

$$M = -\frac{d\bar{a}}{dx}\frac{ART}{GNof}\exp-\frac{zF}{RT}\psi,$$

where M is the unidirectional flux across the membrane, $\bar{a}$ the chemical activity of the species under consideration, A the area of the membrane available for penetration by the species under consideration, G is a frictional coefficient describing the interaction between the membrane and a diffusing ion, No is Avogadro's number, z the valence of the ion, ψ the electrical potential, and f, F, R, and T have their usual meaning.[3] The equation contains several unknowns which are not determinable, and, therefore, it is not too useful in this form. However, if attention is directed to the ratio of fluxes from one side of the membrane to the other, the indeterminate quantities cancel out leaving upon integration the relatively simple relationship,

$$\frac{M_{1.2}}{M_{2.1}} = \frac{f_1C_1}{f_2C_2}\exp\left[\frac{zF}{RT}(\psi_1-\psi_2)\right],$$

which contains only experimentally determinable quantities, except perhaps the activity coefficients, which, however, also cancel out when the solutions on the two sides of the membranes are essentially the same.

The skin was set up between two chambers, each containing solutions vigorously stirred and oxygenated by a stream of air or oxygen. When the inside solution

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

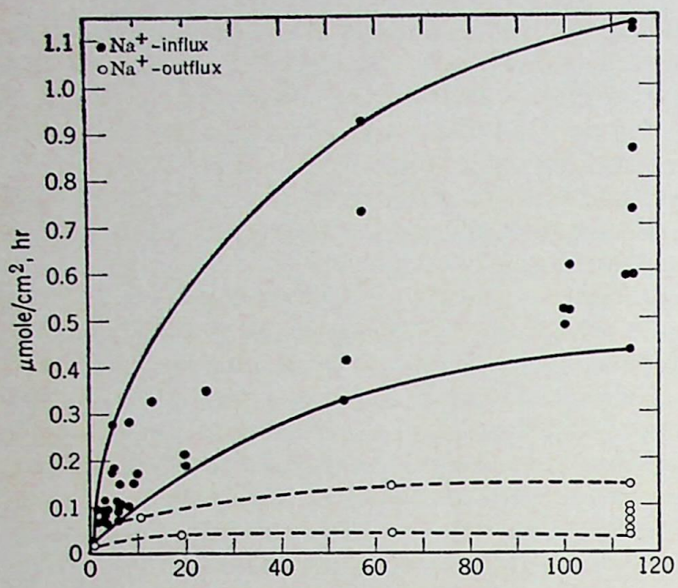


FIG. 1. Sodium fluxes across isolated frog skin as functions of outside sodium concentration [from H. H. Ussing, Cold Spring Harbor Symposia Quant. Biol. 13, 193 (1948)].

was Ringer's solution and the outside various dilutions of Ringer's, it was found[11] that the influx of sodium always was greater than the outflux so long as the outside concentration was not less than about 1 mM/l. The influx of sodium increased in nonlinear fashion as outside sodium was increased; there also was some small increase in outflux but this always remained small (Fig. 1.).

These, of course, represent merely some quantification of what was already well known—that sodium could be transported inward against combined chemical and electrical gradients and was, therefore, an active process. However, it was found that the chloride fluxes corresponded well to the predictions of the flux-ratio equation and that chloride movements could be considered entirely passive.[12] The sodium influx was generally greater than that of chloride so that there could be no question of bulk movements of elements of outside solution to the inside.[13] In general, the potential was higher when the sodium influx was greater and higher when chloride influx was decreased as could be achieved by treating the skin with 10^{-5} MCu^{++} solutions. It was proposed that the potential was produced by the transport of sodium and that the passive movements of other ions tended to reduce the potential.[14]

TABLE I. Unidirectional fluxes, net flux, and electric current in short-circuited frog skin [from H. H. Ussing, *The Relation between Active Ion Transport and Bioelectric Phenomena* (Instituto de Biofisica, Universidade do Brasil, Rio de Janeiro, 1955)].

	μa/cm²			
	Na_{in}	Na_{out}	ΔNa	Current
I	20.1	2.4	17.7	17.8
II	11.1	1.5	9.6	9.9
III	40.1	0.89	39.2	38.6
IV	62.5	2.2	60.3	56.8
V	47.9	2.5	45.4	44.3

Studies of the short-circuited frog skin served to show that only sodium was actively transported and that, when the potential across the skin was kept at zero, there was a remarkable correspondence between the current flowing through the skin and the net movement of sodium ions[14] (Fig. 2 and Table I).

These observations, then, were compatible with the hypothesis that the only active process involved was the transport of sodium ions from outside to inside. No other ion, except—to a small extent—lithium could substitute for sodium,[4] and, in particular, potassium could not, although the complete absence of potassium from the bathing solutions very markedly reduced the transport of sodium—a circumstance to be mentioned again later.

These studies, then, have defined the behavior of the frog skin, but the question remains as to what kind of process might be involved in the transport of sodium

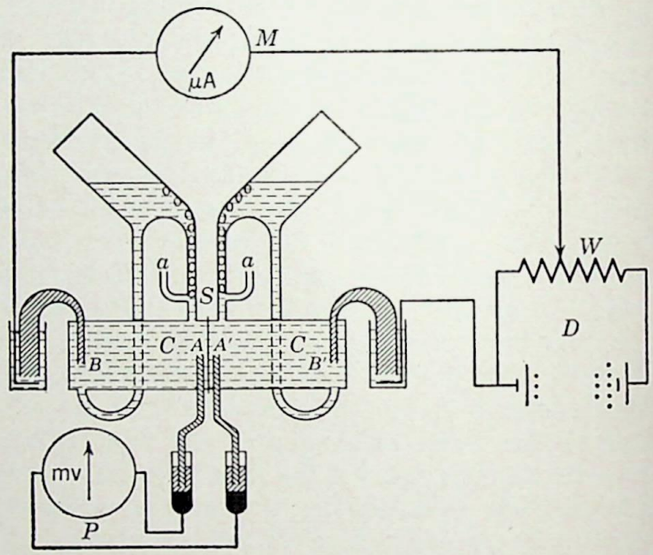


FIG. 2. Apparatus used for determining sodium fluxes and current in short-circuited frog skin [from H. H. Ussing and K. Zerahn, Acta Physiol. Scand. 23, 110 (1951)].

which, of course, is the major question in the field of active transport at present. Two additional types of study involving the amphibian skin have yielded information pertinent to this question.

One of the favorite theories of ion transport has involved the so-called redox or electron-transport pump, variations of which have been suggested by a number of investigators. These hypotheses are based on the assumption that oxidation and reduction of cytochrome iron are spatially separated and that hydrogen ions formed in the reductive step are available for exchange for some cation from outside the cell. The electrons then are passed along the chain, the cations taken up possibly moving with them, and the electrons finally are donated to oxygen to yield hydroxyl ions. Such a scheme has the stoichiometric limitation that only four electrons and four univalent ions can be transported per oxygen molecule consumed. By careful measurement of oxygen consumption and current flow (net

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

sodium transport) in the frog-skin system, both Zerahn[15] and Leaf and Renshaw[16] have been able to show that more than four ions are transported per oxygen consumed. The average figure obtained by Leaf and Renshaw (Fig. 3) was close to seven and some of the ratios were in excess of ten. Thus, even if all of the oxygen consumed were utilized in ion transport—a doubtful assumption—the ratio of transport to oxygen consumption is too great to be attributed to a redox pump mechanism. (Even higher ratios, going up to 20, can be obtained, if one divides the *change* in transport by the *increase* in oxygen consumption when sodium transport is increased in one of several ways.)

An alternative type of mechanism is suggested by other recent observations of Koefoed-Johnsen and Ussing.[5] The experiments were done with membranes in which maximum potentials were induced by reducing to a minimum the effective anion permeability. This

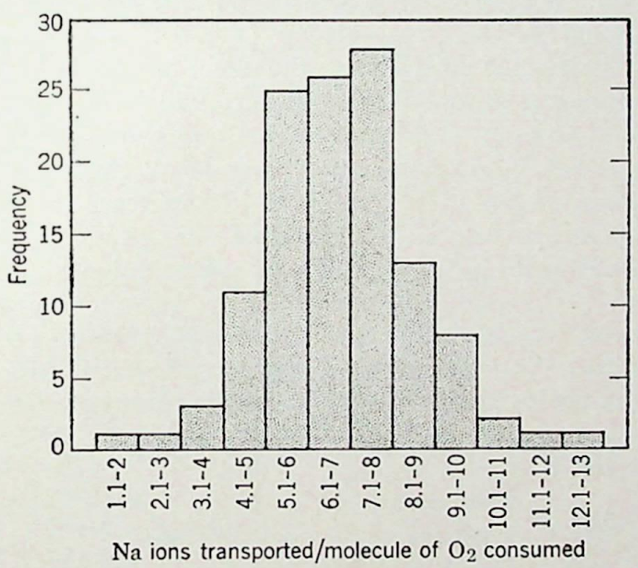


FIG. 3. Relationship of sodium transport to total oxygen consumption of isolated frog skin. Frequency distribution of 120 periods of simultaneous measurement [from A. Leaf and A. Renshaw, Biochem. J. **65**, 82 (1957)].

was done by treating the skin with Cu^{++}, or by replacing the chloride in the outside solution with sulfate to which the skin has a very low permeability. The concentrations of sodium and potassium on each side of the membrane were then varied systematically and the effect on the potential observed. Under these circumstances, the inside of the skin behaved very closely as if it were a potassium electrode, the potential across the skin decreasing approximately 60 mv with each tenfold increase in potassium concentration. Changing the potassium concentration on the outer surface of the skin had no effect, but changing the sodium concentration did, the potential increasing as sodium concentration was increased. This strongly suggested that the two surfaces of the cells making up the skin had quite different and specific properties—a fact that is apparent also from other properties. The two potentials thus behaved as diffusion or junction potentials at boundaries specifically permeable to single cations. The explanation offered is as follows: A sodium ion is extruded from the inner cell surface by forced exchange for potassium; that is, a charged carrier drops off a sodium ion and takes up a potassium ion. At the inner surface, by some exergonic reaction, the specificity of the carrier is changed and it gives up the potassium in exchange for another sodium ion. Thus, the sodium concentration inside the cell is reduced and the potassium concentration raised in a process which, in itself, produces no electrical potential. However, since the inner surface membrane is highly permeable to potassium, potassium diffuses out leading to a diffusion potential with the cell contents negative. The negativity of the cell contents tends to produce the uptake of a cation through the outer surface, but this surface being permeable only to sodium, it is a sodium ion which enters. The net effect is a cycling of potassium at the inner surface and a net movement of sodium ion from outer bathing solution to the inside of the skin. In the presence of chloride and a normal permeability to it, the potential is maintained at a lower level by the diffusion of chloride from outside to inside, and in the normal state the net movement of sodium and chloride, of course, must be very nearly equal.

This model is supported further by potential measurements made by Hoshiko and Engbaek[17] in Ussing's laboratory. By thrusting microelectrodes progressively through the skin, Hoshiko found the change to occur in two jumps which can be interpreted as representing changes occurring at the two surfaces of the active layer of cells.

This, then, provides a well-studied example of electrolyte transport by a membrane and it is generally hoped that, in principle, sodium transport is similar in other systems which are less accessible to isolation and study—such as the electrolyte transport in the renal tubules. Of course, here as elsewhere, very big questions remain—namely, what is the nature of the ion carrier? How does it derive its unusual specificity, and how is this specificity modified to provide the specific orientation of the transport process? Incidentally, mention should be made of the fact that the specificity for sodium on the outward trip (across the inner surface) in frog skin is not matched by a similar specificity for potassium on the inward trip, since varying the *p*H of the inside solution has an effect on the frog skin potential similar to that of changing potassium concentration,[11] suggesting the possibility that hydrogen and potassium may be interchangeable in this process. A similar situation appears to hold in the renal tubule.[18]

So much for the potential-generating and ion-transport properties of frog skin. There remains one aspect not mentioned yet—namely, the roles of diffusion and flow in the movement of water across the skin and the effect of such movements on the movement of solute. These phenomena again can be used to illustrate more-

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

general phenomena characteristic of biological membranes. Actually, even in the conditions already discussed, there is some net movement of water inward in an amount such as to make the net transport of salt result in no net change in solute concentration. However, this is so small as to be justifiably neglected. Under other conditions, however, there may be appreciable net movements of water. How may such water movements be brought about and what might be their effects on the fluxes of water and various solutes? Aside from the possibility of active water transport, a phenomenon existence of which is doubtful and which certainly is not required to explain anything in connection with frog skin, movements of water might be the result of hydrostatic or activity gradients. There are no appreciable hydrostatic gradients involved in the skin system so one is left with gradients of water activity or osmotic pressure. If one has only osmotic gradients, does this exclude the possibility of hydrodynamic flow through pores? It has been claimed that this is the case.[19] However, Ussing has illustrated with a simple model a mechanism by which hydrodynamic flow can result from a gradient of solute concentration only.[13] There are undoubtedly more-rigorous approaches,[20] but this is an easily visualized model. Consider a pore permeating a membrane impermeable to some solute contained in a solution on one side and not on the other. Consider the boundary of the pore at the end facing the solution. Just inside the pore, which the solute cannot enter, is a pure water solution of molar fraction one while just across the boundary of the pore is a solution containing a finite concentration of solute and having a molar fraction of water something less than one. Now clearly there will be a net diffusion of water from just inside the pore into the solution. But since the loss of water from the end of the pore cannot change its molar fraction, there is no reason why water should diffuse from deeper in the pore. Nevertheless, there will be a net movement of water through the pore. This then must be the result of a decrease in pressure at the end of the pore from which water has been lost and, hence, there actually is a hydrostatic gradient causing the water movement and, of course, hydrostatic gradients can cause flow through pores. It has been reported by Mauro that gradients of osmotic pressure across artificial membranes produce flow equal to that produced by the equivalent hydrostatic pressure.[21]

The effect of flow through a pore is to accelerate the diffusion of molecules in the direction of flow and to retard it in the opposite direction. Since the solvent drag force is inversely proportional to the diffusion coefficient, the effect in inducing asymmetry should be greatest for molecules as large as possible which can still pass readily through the pores. Andersen and Ussing have used thiourea and acetamide to examine the phenomenon.[22] Thiourea or acetamide were added to the solution on both sides of the skin (in this case, toad skin because it showed more marked responses to the addition of posterior pituitary extracts) in equal concentration, but the molecules in the outside solution were labelled differently from those in the inside solution. It thus was possible to measure the flux in each direction. Experiments were also done with labeled water. The rate of net movement of water was varied by diluting the outside solution. When the solution on both sides was Ringer's, there was very little net movement of water and the fluxes of the test substance were very nearly symmetrical (Table II). When the outside solution was diluted, there was a net movement of water inward and a marked asymmetry of the fluxes of thiourea and acetamide. When the log of the ratio of inward permeability constant to the outward permeability constant was plotted against the net water transfer, a linear relationship was obtained as predicted theoretically and the ratio of the slopes of 2.3 fits closely with predicted ratios of 2.4 and 2.7 for acetamide and thiourea, respectively (Fig. 4).

Of particular interest was the effect of adding neurohypophyseal hormone (Table II) which greatly increased the net water movements and the asymmetry of the fluxes, but had only a very small effect on the total water flux. Thus, it seemed to increase the area available for flow with very little change in the area available for diffusion. Such an effect might be the result of changing the shape of pores without changing their total area. In any case, these experiments seem to provide a clear indication of the existence of pores and of the production of bulk flow by osmotic gradients.

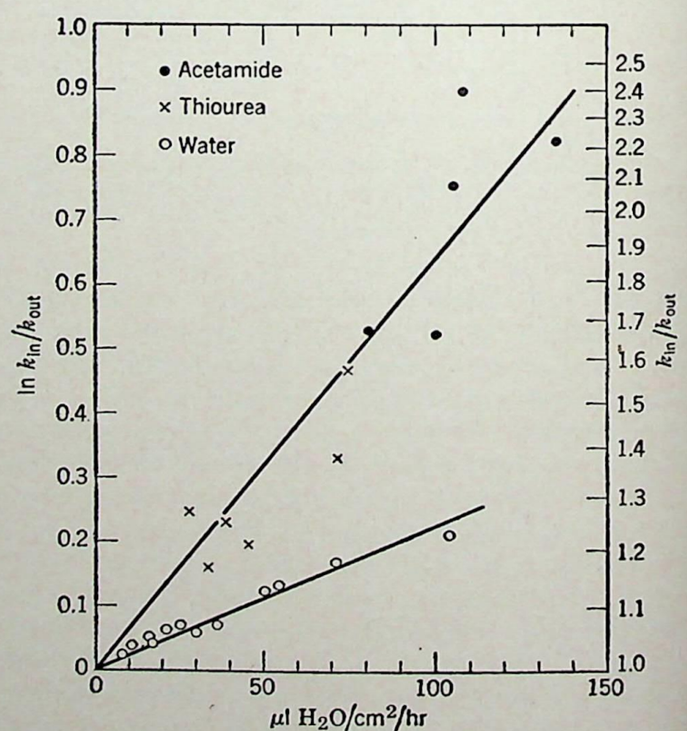


FIG. 4. Relationship between the rate of osmotically induced net water transfer (abscissa) and the logarithm of the ratio of inward to outward rate constants (ordinate) in isolated toad bladder [from B. Andersen and H. H. Ussing, Acta Physiol. Scand. 39, 228 (1957)].

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

TABLE II. Effect of neurohypophyseal hormone on permeability of isolated toad skin. Inside bathing solution Ringer's solution, outside solution as specified. K_{in} = inward permeability coefficient (cm/hr), K_{out} = outward permeability coefficient (cm/hr), and ΔW = net water transfer (μl/cm²/hr), [from B. Andersen and H. H. Ussing, Acta Physiol. Scand. **39**, 228 (1957)].

			Before hormone				After hormone			
Date	Outside medium	Test substance	K_{in} $\times 10^4$	K_{out} $\times 10^4$	$\frac{K_{in}}{K_{out}}$	ΔW	K_{in} $\times 10^4$	K_{out} $\times 10^4$	$\frac{K_{in}}{K_{out}}$	ΔW
March 26	Ringer's	Thiourea	6.63	4.72	1.40	0				
June 11			10.5	8.08	1.30	0	278	278	1.00	0
June 14			11.5	9.75	1.18	0	106	109	0.97	2
June 22			11.5	13.7	0.84	3	247	223	1.11	12
June 23			12.4	14.1	0.88	2	64.7	62.3	1.04	1
June 25			3.17	4.47	0.71	1	7.71	8.45	0.91	1
					1.05	*			0.98	*
June 15		Acetamide					537	516	1.04	14
June 22							363	321	1.13	10
March 25	1/10	Thiourea	2.90	3.03	0.96	8	20.6	16.2	1.27	27
April 6	Ringer's		6.48	6.51	1.00	14	17.4	14.4	1.21	45
April 8			11.4	8.73	1.33	14	49.3	31.2	1.58	74
April 26			4.00	4.02	1.00	8	7.52	6.42	1.17	33
April 29			4.25	4.07	1.04	17	15.4	12.3	1.25	38
May 3			9.86	6.52	1.52	14	156	113	1.38	71
May 25		Acetamide					251	121	2.08	105
June 1							175	105	1.67	80
June 6							299	180	1.66	100
June 10							530	239	2.22	135
June 29			129	100	1.29	24				
April 6			81.5	76.1	1.16	16	196	81.5	2.40	108
April 26		Heavy water	3730	3510	1.06	22	5550	4510	1.23	104
May 3			4090	3890	1.05	20	4250	3750	1.13	50
May 9			3630	3420	1.06	21	4270	3730	1.14	54
May 16			4040	3760	1.07	28	4560	3850	1.18	71

They also make unnecessary any assumed active water transport by the amphibian skin. Similar phenomena also have been demonstrated to occur in some marine eggs,[23] and it is a reasonable conclusion that the earlier observation of unexplained high net water fluxes in intestinal loops has a similar basis.

The phenomenon of exchange diffusion is of interest with respect to the mechanism by which ions cross cell membranes. The concept was invented by Ussing[11] to rationalize the very high rate of isotopic exchange between the sodium in cells with that in their surroundings. Thus, with the potential inside negative relative to the outside and with the sodium concentration much higher outside, work is required to extrude sodium. If it is assumed that the inward movements are entirely by diffusion and that the outward movement proceeds at the expense of metabolic energy, then the energy requirements would exceed half of the total metabolic energy production of muscle in the resting state. Ussing proposed that sodium ions attached to a charged carrier might exchange for other sodium ions in the external solution or in the cell interior, the carrier making the round trip with sodium attached so that no net change in sodium concentration on either side of the membrane would occur. Yet, if the sodium on one side of the membrane were isotopically labeled, the process would result in movement of label.

The concept of exchange diffusion has been useful in the interpretation of a number of other phenomena some of which more directly are evidence for the idea of reversible combination of the same ion with a carrier so that exchange is greater than actual transport or movement of the free ion. For instance, Hogben[24] has studied the secretory membrane of the frog stomach using the Ussing technique. The stomach secretes hydrochloric acid, chloride against both electrical and concentration gradient, hydrogen ion with the electrical gradient which, however, is smaller than the very steep opposing concentration gradient. Sodium behaves passively. However, the striking phenomenon in connection with the phenomenon of exchange diffusion is the very large flux in the direction opposite to active transport (Table III). The flux from the secretory surface corresponding to the inside of the stomach, to the nutrient surface corresponding to the blood, is more than twice the membrane conductance as determined by the relationship between imposed current and developed potential. A considerable part of the chloride

TABLE III. Chloride fluxes in stomach mucosa of the bull frog [from A. Hogben, *Electrolytes in Biological Systems*, A. Shanes, editor (American Physiological Society, Washington, 1955), p. 176].

Net chloride transfer μ Eq cm^{-2} hr^{-1}		Net charge transfer μ Eq cm^{-2} hr^{-1}	
Flux N−S	10.65	Current	3.05
Flux S−N	6.38	H^+ secretion	1.20
	4.27		4.25
Mean conductivity:	Start 3.21 m mhos		
	End 2.28 m mhos		

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

TABLE IV. Na active transport and exchange diffusion in HK sheep red cells.

External cation	Na outflux	Difference
Na+K	3.5	2.7 E.D.[a]
Mg+K	0.78	
Mg+K+Stroph	0.16	0.62 A.T.[b]
Mg	0.16	

[a] E.D. = Exchange diffusion.
[b] A.T. = Active transport.

movement must occur in a form which does not contribute to the transfer of charge—presumably the same carrier complex by which the active transport is effected.

Another similar case involves the movement of chloride across the large intestine of the frog and has been studied by Cooperstein and Hogben.[25] In this case, there is no active transport, chloride fluxes being symmetrical when the electrical potential is zero. However, at various levels of potential, the flux ratio is lower than predicted from the electrochemical potential ratio. This could be explained if some part of the flux were owing to exchange diffusion and thus was not affected by the electrochemical potential gradient.

Finally, an example of quite a different sort involving sodium transport by the red blood cell. The red cell is rather unusual, in a number of respects, in that it is non-nucleated, lacks the machinery for oxidative metabolism, and, in general, transports sodium and potassium at rates several orders of magnitude lower than most body cells. Also, they contain a rather high chloride concentration, 60 to 70% of that in the plasma, whereas most other cells contain very little; since they have an extremely high permeability to chloride, with a half-turnover time, measured in milliseconds,[26] the presumption is that the distribution is passive and, therefore, that the electrical potential can be estimated from the chloride-distribution ratio. If this is true, the accumulation of potassium in this cell cannot be considered as a passive consequence of sodium extrusion, as is frequently considered the case in other cells. In any case, the slow rate of transfer and the ease of direct analysis make the red cell a suitable object for study and the relative simplicity of the red-cell metabolism somewhat facilitates the study of the metabolic correlates of ion transport.

Some data obtained with sheep red cells by Tosteson and Hoffman[27] in our laboratory illustrate the phenomenon of exchange diffusion (Table IV). When cells are incubated in normal plasma, an outflux of sodium of 3.5 mM/l/hr is observed. The removal of sodium and its replacement with magnesium very markedly depresses the outflux of sodium. Since the only type of process by which sodium leaves cells which we can visualize as requiring sodium on the *outside* is exchange diffusion, it appears that close to 80% of the sodium output from the sheep red cell is attributable to this process. When the potassium also is removed from the medium, sodium output is reduced even further to a level compatible with complete absence of active transport and which is the same low level reached when transport is inhibited with strophanthanthidin, a glycoside related to digitalis and a member of a group of powerful and apparently specific inhibitors of sodium transport. These results again suggest that the sodium extrusion is effected by a carrier-linked exchange for potassium. From the amount of glycoside required to inhibit potassium uptake by human red blood cells, Glynn[28] has estimated that there are less than 1000 transport sites on the surface of the human red blood cell, a disk 7 μ in diameter and perhaps 1 μ thick.

These similarities between membrane-transport phenomena in diverse systems and even in orders of vertebrates encourage one to believe that the processes involved fundamentally are limited in number and are not different in each cell type and species, and that the study of simpler systems may yield information applicable to more-complex and less-accessible ones.

BIBLIOGRAPHY

[1] E. Overton, Vierteljahresschr. Naturforsch. Ges. Zürich **40**, 159 (1895).

[2] J. R. Pappenheimer, E. M. Renkin, and L. M. Borrero, Am. J. Physiol. **167**, 13 (1951).

[3] H. H. Ussing, Acta Physiol. Scand. **19**, 43 (1949).

[4] H. H. Ussing, Symp. Soc. Exptl. Biol. **8**, 407 (1954).

[5] V. Koefoed-Johnsen and H. H. Ussing, Acta Physiol. Scand. **42**, 298 (1958).

[6] A. Leaf, J. Anderson, and L. B. Page, J. Gen. Physiol. **41**, 657 (1958).

[7] A. Krogh, Skand. Arch. Physiol. **76**, 60 (1937).

[8] G. von Hevesy, E. Hofer, and A. Krogh, Skand. Arch. Physiol. **72**, 199 (1935).

[9] M. B. Visscher, E. S. Fetcher, Jr., C. W. Carr, H. P. Gregor, M. S. Bushey, and D. E. Barker, Am. J. Physiol. **142**, 550 (1944).

[10] F. Brunn, Z. ges. exptl. Med. **25**, 170 (1921).

[11] H. H. Ussing, Cold Spring Harbor Symposia Quant. Biol. **13**, 193 (1948).

[12] V. Koefoed-Johnsen, H. Levi, and H. H. Ussing, Acta Physiol. Scand. **25**, 150 (1952).

[13] H. H. Ussing, *The Relation between Active Ion Transport and Bioelectric Phenomena* (Instituto de Biofisica, Universidade do Brasil, Rio de Janeiro, 1955).

[14] H. H. Ussing and K. Zerahn, Acta Physiol. Scand. **23**, 110 (1951).

[15] K. Zerahn, Acta Physiol. Scand. **36**, 300 (1956).

[16] A. Leaf and A. Renshaw, Biochem. J. **65**, 82 (1957).

[17] T. Hoshiko and L. Engbaek, Abstr. Communications of the 20th International Physiological Congress, Brussels (1956), p. 443.

[18] R. W. Berliner, T. J. Kennedy, Jr., and J. Orloff, Arch. Int. Pharmacodynamie **97**, 299 (1954).

[19] F. P. Chinard, Am. J. Physiol. **171**, 578 (1952).

[20] O. Kedem and A. Katchalsky, Biochim. et Biophys. Acta **27**, 229 (1958).

[21] A. Mauro, Science **126**, 252 (1957).

[22] B. Andersen and H. H. Ussing, Acta Physiol. Scand. **39**, 228 (1957).

[23] D. M. Prescott and E. Zeuthen, Acta Physiol. Scand. **28**, 77 (1953).

[24] A. M. Hogben in *Electrolytes in Biological Systems*, A. Shanes, editor (American Physiological Society, Washington, D. C., 1955), p. 176.

[25] I. L. Cooperstein and C. A. M. Hogben, Federation Proc. **17**, 29 (1958).

[26] D. C. Tosteson, Abstr. Communications of the 20th International Physiological Congress, Brussels (1956), p. 892.

[27] D. C. Tosteson and J. Hoffman (personal communication).

[28] I. M. Glynn, J. Physiol. **136**, 148 (1957).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

REVIEWS OF MODERN PHYSICS VOLUME 31, NUMBER 2 APRIL, 1959

41

Interaction Properties of Elongate Protein Macromolecules with Particular Reference to Collagen (Tropocollagen)*

FRANCIS O. SCHMITT

Department of Biology, Massachusetts Institute of Technology, Cambridge 39, Massachusetts

ORGANIZATION and the interaction of organized structures at all levels of complexity are of cardinal significance in life processes. This organization may be manifested within the protein or other biomolecule or may depend upon the interaction of many individual entities.

For purposes of biophysical and biochemical investigation of such interaction of organized systems, it is profitable to deal with relatively simple systems, preferably with pure substances. Such systems are available in materials which are primarily one-dimensional (fibrous), two-dimensional (membranous), or three-dimensional (crystalline). The present paper is concerned exclusively with the properties of highly elongate, fibrous macromolecular systems. In a study of the forces between such highly organized fibrous systems lies rich opportunity for fundamental biophysical research.

Many substances of crucial importance in cells and tissues occur as very thin (10 to 30 A), highly elongate (1000 to 5000 A) particles. Proteins (such as myosin, paramyosin, actin, fibrinogen-fibrin, collagen, and the nerve-axon protein), nucleic acids (DNA and RNA), and polysaccharides (such as cellulose, hyaluronic acid, chondroitin sulfate) are examples. Many of these substances are themselves polymers (as the protein macromolecules are polymers of many amino-acid residues) but, as monomers, these elongate macromolecules polymerize end-to-end and laterally to form fibrous structures.

These macromolecular systems lend themselves to detailed physicochemical, crystallographic, and electron-optical study. Polyelectrolyte theory may be applied fruitfully to them. In many cases, energy is coupled with the macromolecular system by interaction with "energy-rich" substances such as adenosine triphosphate (ATP), but the mechanism by which this available energy is caused to do work (mechanical, chemical, electrical, or osmotic) is still poorly understood.

SOME TYPICAL BIOLOGICAL SYSTEMS IN WHICH FUNCTION DEPENDS UPON SPECIFIC INTERACTION OF ELONGATE MACROMOLECULES

It will help to identify the types of biophysical problems involved by mentioning a few typical examples.

1. Functions Involving Mechanical Properties

In this category, mention is made of but three types that may serve to illustrate fundamental problems: fibrous systems designed to afford high tensile strength; those which produce tension or undergo contraction; and those which, by forming gelled clots, occlude regions such as blood vessels and prevent bleeding.

An excellent example of a fibrous system that achieves high tensile strength (*ca* 100 000 lb/in.^{2}) by a lateral bonding of macromolecular polymers are the structures such as skin, tendon, and other forms of connective tissue involving the protein collagen. The macromolecules are synthesized within fibroblast cells, find their way into the intercellular tissue spaces, and eventually aggregate at the appropriate place and time to form fibers. The mechanism of fibrogenesis may be very complex, involving processes of activation and homeostatic control of such processes so as to facilitate fibrogenesis when needed (as in wound repair) and to prevent excessive fibrinogenesis (such as occurs in aging and in certain pathological processes as in atherosclerosis, and in certain rheumatoid and so-called "collagen diseases").

The second mechanical function, that of contraction or tension production, poses a problem as to whether the shortening occurs essentially as an intramolecular process of superfolding of polypeptide chains, or is rather an intermolecular process involving a rapid and reversible change in the affinity or interaction between two or more species of fibrous proteins leading to a shortening of the fibrous system without substantial change in the helical configuration of the intramolecular chains characteristic of the native macromolecules.

Another rather striking problem is well illustrated in the embryogenesis of fibrous structures such as striated muscle, in which the axial repeat (sarcomere length) is so precise as to give several orders of diffraction with visible light. This production of supermacromolecular patterns well may involve the specific aggregation of several species of macromolecules, each having lengths of the order of several thousands of Ångström units. Perhaps some day it will be possible to produce such super-repeating patterns by interacting several kinds of fibrous macromolecules (protein, nucleic acid, or polysaccharide) under appropriate conditions *in vitro*.

The third type of mechanical function mentioned in the foregoing is that of the clotting of fibrous protein,

* These studies were aided by a research grant from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Public Health Service, U. S. Department of Health, Education, and Welfare.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

as in the transformation of the soluble fibrinogen in the blood into the insoluble fibrin of blood clot. This involves an enzymatic activation of the soluble protein monomers by means of an enzyme (itself activated by a series of interdependent processes) to produce "intermediate polymers" several thousand Ångström units long which then polymerize spontaneously to form the clot. The enzymatic activation itself is controlled by a highly complex system of kinases and antikinases by means of which a high degree of homeostatic control of this vital process is achieved.

2. Functions Involving Enzyme Action

As was emphasized by Engelhardt and Ljubimova,[1] large macromolecules may function enzymatically. This may occur either because the macromolecule as a whole acts like an enzyme or because a portion of the molecule is enzymatically active. Myosin, with which Engelhardt and Ljubimova were concerned,[1] was found to exert enzymatic action in splitting ATP. It is known now that only one component ("heavy meromyosin") is enzymatically active as ATPase. It remains to be seen how many other kinds of elongate macromolecules will be found to have enzymatic properties. It may be mentioned that, when an enzymatic group forms part of a large macromolecular complex, the configuration of the macromolecules or smaller molecules in the environment, for steric reasons, may strongly influence the availability of the enzymatic site. Such regulatory action well may be involved in muscle contraction.

3. Functions Involving the Maintenance of a Specific Linear Sequence of Chemical Groups as in Genetic Determiners

It is believed that in the linear sequence of nucleotide residues in the double helix of DNA is to be found the coding responsible for transmitting genetic information. The DNA occurs as highly elongate macromolecules extractable from the chromosomes by mild methods (such as by treatment with hydrogen bond breakers). It seems probable that the ability of such macromolecules to exert their specific controlling influence during development and differentiation must depend importantly upon their interaction with other constituents of the chromosomes such as histones, protamines, and protein macromolecules, and perhaps also with other nucleic-acid macromolecules.

4. Other Functions

Engelhardt[2] suggested that macromolecules may perform other types of work such as osmotic and electrical. These possibilities have not yet been explored thoroughly. The function of certain fibrous proteins, such as the axoplasmic protein of nerve, remains completely unknown.

Perhaps one may mention in this connection also the action postulated by Weiss[3] as important in determining the ordering of cells into typical tissues by means of the interaction of specific types of macromolecules at the surfaces of the cells that form the tissues. Presumably because of a complementarity or ordered type of interaction of the surface molecules, the cells are caused to aggregate in patterns characteristic of each tissue.

From this very brief description, it is obvious that crucial biological functions depend importantly upon the ways in which highly specifically structured macromolecules interact with one another and upon the manner in which environmental conditions change or regulate this interaction. It is the purpose of this paper to illustrate this specificity of interaction by a consideration of the properties of one particular type of macromolecule, collagen, chosen because of its specially favorable chemical and structural properties. Like many other biological macromolecules, collagen is very asymmetric (*ca* 14×2800 A). It has, however, the valuable property that, in its aggregation patterns, it forms fibrous structures characterized by highly specific band patterns as seen in the electron microscope. By analysis of these band patterns, it is possible to deduce the type of macromolecular interaction responsible for each characteristic pattern. From such studies, valuable lessons are learned that have direct application to other types of macromolecular systems in which no band patterns exist to serve as guides.

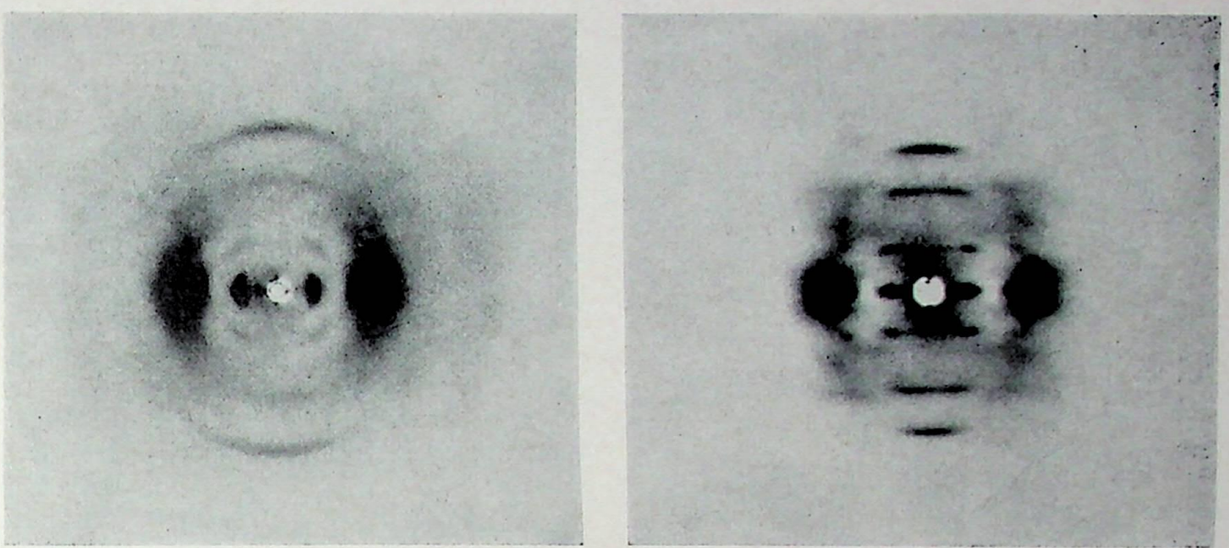
BIOPHYSICAL AND BIOCHEMICAL PROPERTIES OF COLLAGEN

In contradistinction to most proteins, collagen (or perhaps more correctly the collagen class) possesses characteristic structural and chemical properties which permit its definitive identification. For present purposes, it is necessary only to sketch those properties necessary for an understanding of the internal structure, and the chemical properties of the collagen macromolecule. For recent excellent surveys of these properties, see Gustavson,[4] Highberger,[5] the CIOMS Symposium on Connective Tissue,[6] and that on Gelatin and Glue Research.[7]

Collagen occurs in dense fibrous tissue of high tensile strength, as in tendons, or less tightly woven tissue fabrics, as in skin, or in more sparse distribution as in loose connective tissue. The fibrous protein occurs in various hierarchies of fiber size, including the following: *fibers*, visible macroscopically or microscopically and having diameter of the order of micra; *fibrils* with widths of the order of a few hundred to several thousand Ångström units, observable in the dark-field microscope and resolvable in the electron microscope; the *protofibrils* which originally were defined as constituting "The unit colúmnar arrays which, when associated laterally, form the collagen fibril";[8] and the *collagen* (or "tropocollagen") *macromolecules* which constitute the monomeric units of the protofibrillar polymer.

For x-ray diffraction and for chemical-analytical

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA



(a) (b)

FIG. 1. Large-angle x-ray diffraction patterns of collagen from rat-tail tendon.[10] (a) Unstretched; (b) stretched 8% [from J. T. Randall, J. Soc. Leather Trades' Chemists 38, 362 (1954)].

studies, gross macroscopic fibers or whole tissues are used. For electron-microscopic investigation, the fibrous category of interest is the fibril which manifests a detailed and characteristic band pattern which, as is brought out below, results from a specific pattern of aggregation of the elongate native collagen macromolecules.

The collagen class of proteins, as Astbury[9] referred to them, is uniquely characterized by its amino-acid composition, its x-ray diffraction pattern, and its banded appearance in the electron microscope. These characteristics may be described briefly as follows.

The collagenous proteins differ from other proteins in that they contain the amino acids hydroxyproline and hydroxylysine. In mammalian collagen, about one-third of the amino-acid residues are glycine. Proline and hydroxyproline together make up almost another third, leaving approximately one-third for other amino-acid types. From a determination of the hydroxyproline and glycine content of a given preparation, one can make a good estimate of the collagen content.

Perhaps the most distinctive characteristic of collagen is its large-angle x-ray diffraction pattern (Fig. 1) which reflects the internal organization of the collagen macromolecule and is, therefore, characteristic of this class of proteins. Astbury[9] early called attention to the 2.86-A meridional reflection, which he considered to represent the length of the amino-acid residue, along the fiber axis, in a coiled polypeptide chain, and to the equatorial reflections at 10 to 15 A (depending upon the degree of hydration) which he attributed to the separation between main chains. With more refined technique, it has been possible to obtain far more reflections in the pattern and to achieve a higher degree of orientation by stretching fresh tendon. From such patterns, it has been possible for several groups of workers to agree that the diffractions are interpreted best in terms of a macromolecule containing three chains coiled in helical fashion about each other to form a coiled coil (see particularly the papers of Crick and Rich).[11–13] The proposed triple-stranded structure is shown schematically in Fig. 2.

It has been proposed also that there are only two types of three-stranded helical models of collagen structure, based on the so-called structure I and structure II, derived from a consideration of polyglycine, and compatible with the x-ray, infrared, analytical, and physicochemical data. In these models, the axial repeat occurs at 28.6 A. In collagen II, to agree best with the diffraction data, the OH groups of the hydroxyproline residues extend radially from the three chains, making it possible to form hydrogen bonds with CO groups of adjacent three-stranded macromolecules. In the collagen I structure, the hydrogen bonds from hydroxyproline are directed internally, bonding the three chains intramolecularly. The hydroxyproline content appears to be determinative of the denaturation temperature, which is a measure of the energy needed to disrupt the internal organization of the macromolecule. This fact tends to support the collagen-II type of structure. Rich (p. 50) has suggested that one type of structure might be convertible into the other and that this may result from application of stress to the fiber.

Although there is fairly general agreement that the collagen macromolecule is a three-stranded helix, it is not certain that the macromolecule is thus constructed over its entire extent. Gallop[14] suggests that as much as 30% may have a different configuration.

Treatment of soluble collagens with hydrogen-bond breakers like urea, or by heating, causes denaturation with the liberation of the constituent chains to form "parent gelatin." From the original macromolecule, having a weight of 360 000, there is formed, according to Doty and Nishihara,[15] one chain with a weight of

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

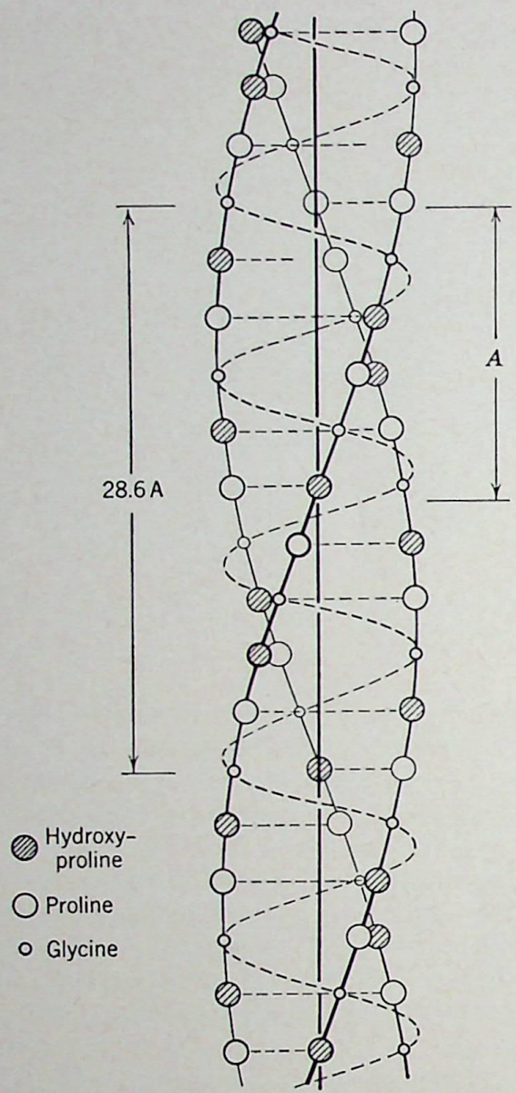


FIG. 2. Three-stranded helical structure of collagen macromolecules (courtesy A. Rich).

120 000 and another with a weight of 240 000. These authors believe that an alkali-labile ester bond links two chains of weight 120 000 to form the heavier chain obtained from denatured collagen. This is to be compared with the corresponding values of Orekhovitch and Shpikiter.[16]

In addition to the large-angle x-ray pattern, arising from the internal, presumably three-stranded, structure of the macromolecule, collagen also manifests a well-developed small-angle x-ray pattern (Fig. 3) consisting of many (*ca* 50) orders of a large axial repeat which Bear[18,19] showed to be 640 A in air-dried fibers and nearer to 700 A in moist fibers. Although all native collagen fibers from a wide variety of sources showed this axial repeat, its significance in terms of molecular structure was not obvious. The simplest early interpretation was that it represents the molecular length of the collagen molecules, an assumption that seemed to gain support from the fact that a similar axial repeat was observed in the band pattern observed in electron micrographs. Bear and Morgan[20] attempted to relate the positions of the intraperiod bonds observed electron optically with the characteristic intensities of the various orders of the small-angle x-ray pattern. As is shown in the following, the collagen macromolecule probably has a length of four times the 700-A period, i.e., about 2800 A.

The band pattern observed in high-resolution electron micrographs of teased collagen fibrils stained with phosphotungstic acid (PTA), or other heteropolyacid, is uniquely characteristic of collagen (see Fig. 4). This axial pattern repeats at about 700 A and contains a number of bands and interbands of characteristic density and position. It was suggested by Bear[21] that the bands represent regions of relative disorder due to the interaction of side chains of relatively large size while the interbands represent regions of relative order due to the interaction of the smaller side chains which are found in considerable abundance in collagen. Another interpretation of band structure depends upon the characteristic interaction of groups such as the guanidino groups of the arginine side chains with PTA, as suggested by Kühn *et al.*[22] It is noted from Figs. 4 and 5 that the band pattern—i.e., the intraperiod positions and the relative densities of the bands—is a polarized, asymmetric pattern. The significance of this pattern was discovered only after it became possible to take the native fibrils apart into their constituent macromolecules and to cause these to re-aggregate in characteristic and new band patterns. These results may now be described briefly.

FORMATION OF ORDERED AGGREGATION STATES OF COLLAGEN BY PRECIPITATION FROM SOLUTION

A formidable difficulty in the characterization of the collagen molecules lay in the relative insolubility of collagen fibers. However, certain types of collagen, such as in rat-tail tendon and in the fish swim bladder, are

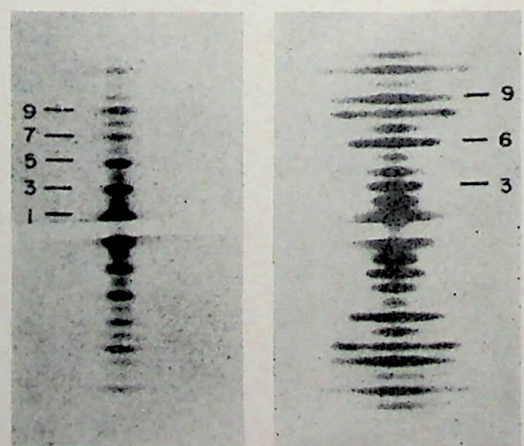


FIG. 3. Small-angle x-ray diffraction patterns of kangaroo-tail tendon collagen; (a) moist preparation; (b) after brief exposure to water and drying under tension.[17] Layer line indices are indicated [from R. S. Bear, O. E. A. Bolduan, and T. P. Salo, J. Am. Leather Chemists' Assoc. **46**, 107 (1951)].

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

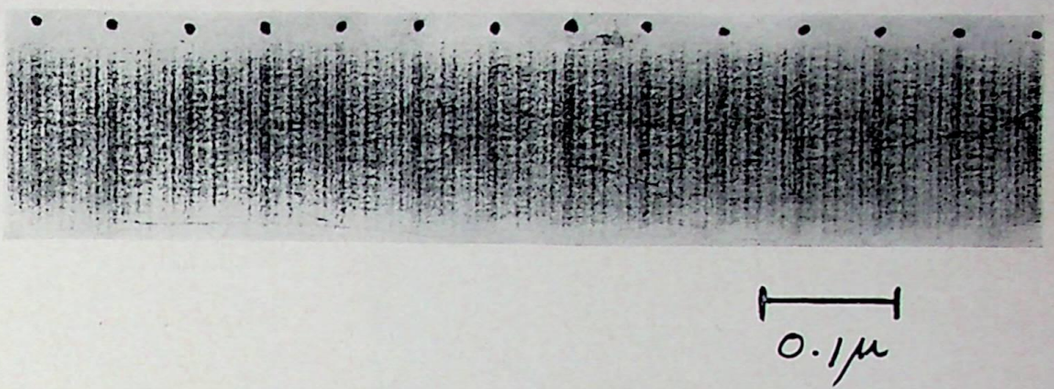


FIG. 4. Electron micrograph of calfskin collagen reconstituted from solution, stained with phosphotungstic acid. Band pattern is of the native type. Axial repeats marked (courtesy A. J. Hodge).

soluble in dilute acid. From the classical early work of Zachariades, Nageotte, Fauré-Fremiet, Wyckoff, and Corey, and others, it is known that, by appropriate adjustment of the *p*H and ionic strength of such acid solutions, the collagen can be precipitated reversibly in fibrous form. Examined in the electron microscope after staining with PTA, the reprecipitated fibrils were found to possess structure, the type of which depends upon the conditions of precipitation. With increasing ionic strength, the band pattern may be that characteristic of native fibrils (period ≅700 A), it may be about one-third this value, or the precipitate may have tactoidal appearance, showing no bands at all. These different forms can be produced reversibly from acid solutions of highly purified collagen; presumably the different ordered states depend only upon the collagen and require no additional organic material.

When certain types of extracts are made from connective tissue or when certain organic substances, particularly highly negatively charged substances, are added to the collagen solutions and the conditions are adjusted appropriately, a new modification is found which manifests an axial repeat or identity period about four times that of normal collagen (i.e., about 2600 to 3000 A) and which, therefore, are called "long-spacing" types. Two such forms, called "fibrous long-spacing" (FLS) and "segment long-spacing" (SLS) are shown diagrammatically in Fig. 5. The FLS modification is produced routinely by addition of α-1 acid glycoprotein to an acetic-acid solution of collagen, followed by dialysis against water. The SLS modification is produced routinely by addition of ATP to the acid solution of collagen; the precipitate forms directly without further adjustment of conditions.

It is noted that the FLS type has a symmetrical, nonpolarized band structure while the SLS has an asymmetrical, polarized pattern of banding.

Each of the five band patterns described may be produced reversibly from an acid solution of collagen. The particular patterns produced depend for their specificity upon the collagen rather than upon the other substances added or conditions imposed; rather these substances and conditions serve to evoke the structure inherently characteristic of the collagen itself.

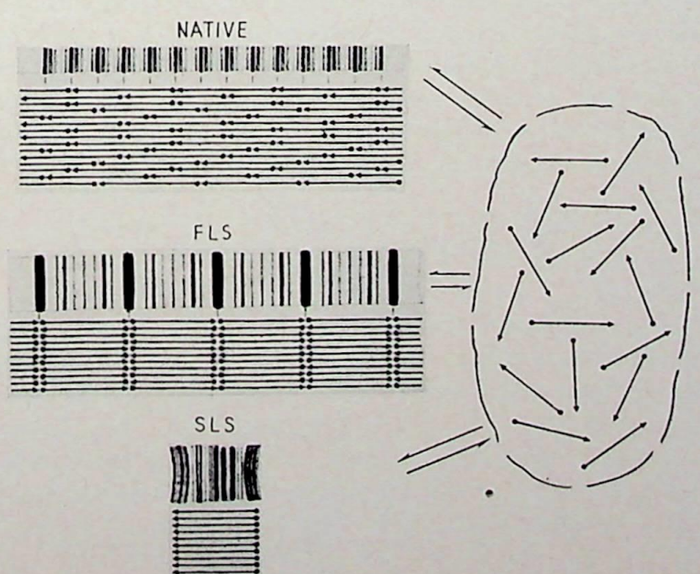


FIG. 5. Diagrammatic illustration of patterns of aggregation of tropocollagen macromolecules in native, FLS, and SLS types. Polarization of macromolecules indicated by arrow.

MACROMOLECULAR MONOMER OF COLLAGEN—THE "TROPOCOLLAGEN" HYPOTHESIS

The structures described in the foregoing, discovered in collaboration with Gross and Highberger, were interpreted as follows (see summaries of this work by Schmitt, Gross, and Highberger;[23] Schmitt;[24] Gross;[25] and Highberger.[5] It is assumed that the long-spacing (about 2800 A) represents the length of the native collagen macromolecule which has a three-stranded helical internal structure, as deduced from the large-angle x-ray pattern (see the foregoing). The long, thin macromolecules were given the term "tropocollagen" (TC), because they are capable of "turning into" or forming the native collagen structure, and also to distinguish them from various other collagen fractions (such as procollagen) previously described.

The TC macromolecules are assumed to be essentially identical in structure and composition and to be themselves polarized in the sense of the linear sequence of amino-acid residues in the constituent intramolecular strands.† This is indicated by the arrows on the TC

† The unit of native collagen structure is referred to as a macromolecule rather than as a molecule because it appears to be composed of several covalent polypeptide chains bonded together by hydrogen bonds.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

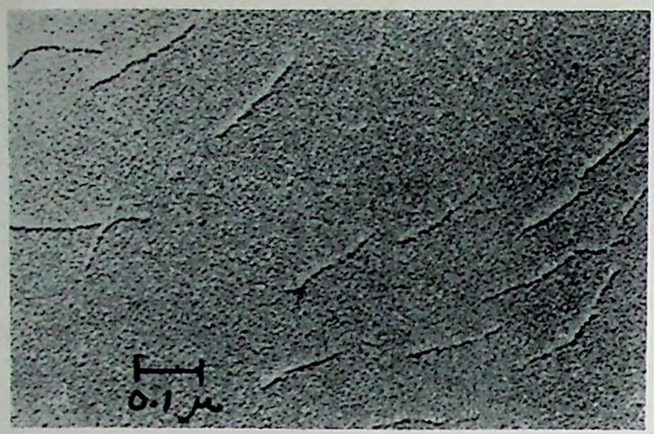


FIG. 6. Electron micrograph of tropocollagen macromolecules prepared by the method of Hall[28,29] (courtesy C. E. Hall).

macromolecules in Fig. 5. The hypothesis assumes that the various types of ordered patterns of TC aggregation occur by virtue of relatively stable bonding between terminal groups on the side chains of laterally adjacent macromolecules. Each type of ordered aggregation type represents a particular pattern of interacting side chains. In the SLS form, it is assumed that the TC macromolecules are essentially in register with respect to their ends and are "pointing" all in the same direction; i.e., they are in parallel array. The SLS pattern, therefore, provides a molecular "fingerprint" of the sequence of amino-acid residue types along the TC macromolecule—information not deducible by examination of the band pattern of native fibrils. The FLS is assumed to be formed by an antiparallel packing of TC in which the macromolecular ends are approximately in register (see Fig. 5).

Since the axial repeating pattern of native fibrils, both as seen in electron micrographs and as measured in the small-angle x-ray pattern, is about a quarter of that of the length of the TC macromolecules, it was assumed that the latter are arranged in parallel array but are displaced in the axial direction by one-quarter of a length in adjacent macromolecules (see Fig. 5). A specific suggestion somewhat along the same line has been proposed by Tomlin and Worthington.[26]

This concept of the structure and properties of the native macromolecule of collagen was deduced from the electron-optical observations of the various ordered aggregation types observed. The hypothesis received confirmation from the physicochemical studies by Boedtker and Doty[27] performed on solutions which were highly monodisperse with respect to the monomer macromolecules (achieved by centrifuging out the larger aggregates). These data indicated that the macromolecules behave like rigid rods with dimensions about 14×2800 A and molecular weight about 360 000. Previous estimates of other workers about particle sizes were considerably greater, probably because their preparations were heterodisperse, containing polymers of collagen as well as the monomers.

Finally, the tropocollagen macromolecules were visualized directly in the electron microscope by a method developed by Hall.[28] This consists in depositing the molecules upon the atomically smooth surface of freshly cleaned mica by spraying a very dilute solution of the protein. After drying, this surface is shadowed by evaporation of platinum at a small angle. The metalized layer then is backed with a thin collodion supporting film, stripped from the mica, and examined in the electron microscope at high resolution. From such electron micrographs (see Fig. 6), Hall[29] found the fibrous particles to be about 15 A in width but their lengths to be somewhat smaller than had been predicted by the physicochemical data of Boedtker and Doty. Subsequently, with improved technique, Hall and Doty[30] obtained a weight-average length of 2820 A, in good agreement with the physicochemical data and with the lengths determined in this laboratory on the same solutions used by Hall and Doty by conversion to the FLS modifications and measurement of the axial period (average value was 2700 A).

The problem of the nature of the precursor of fibrous collagen in the fibrils of connective tissue has been the subject of much investigation. Orekhovitch *et al.*[31] suggested that the fraction soluble in citrate buffer ($\mu \cong 0.2$, $p\mathrm{H} \cong 3.5$) is such a precursor and, therefore, gave the material the name "procollagen." However, from turnover studies or the incorporation of C^{14}-labeled glycine, Harkness *et al.*[32] suggested that the precursor is to be found in the material soluble in slightly alkaline buffer, with a much shorter half-life than citrate-soluble collagen. This conclusion was confirmed by Jackson[33] using other methods. The possibility that tropocollagen macromolecules soluble in neutral salt solutions (Gross, Highberger, and Schmitt,[34]) may be the precursor of fibrous collagen has been discussed in some detail by Gross.[35] Orekhovitch and Shpikiter[16] have concluded that procollagen and tropocollagen are, in fact, identical. The possibility that collagen, as synthesized in the fibroblasts, requires activation before it is capable of being incorporated into fibrous tissue has been much investigated, but the details of the process remain to be disclosed.

FRAGMENTATION OF TROPOCOLLAGEN MACROMOLECULES BY SONIC IRRADIATION

The discovery by Nishihara and Doty[36] that sonic irradiation of tropocollagen rapidly reduces the viscosity without substantial reduction in optical rotation suggested that the irradiation fragments the macromolecules into shorter pieces, actually into halves and quarters, which retain the triple-chain helical structure characteristic of the native macromolecules. This possibility was confirmed by Hodge and Schmitt[37] by electron-microscopic examination of irradiated collagen. The loci along the macromolecules which undergo scission could be determined with considerable precision

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

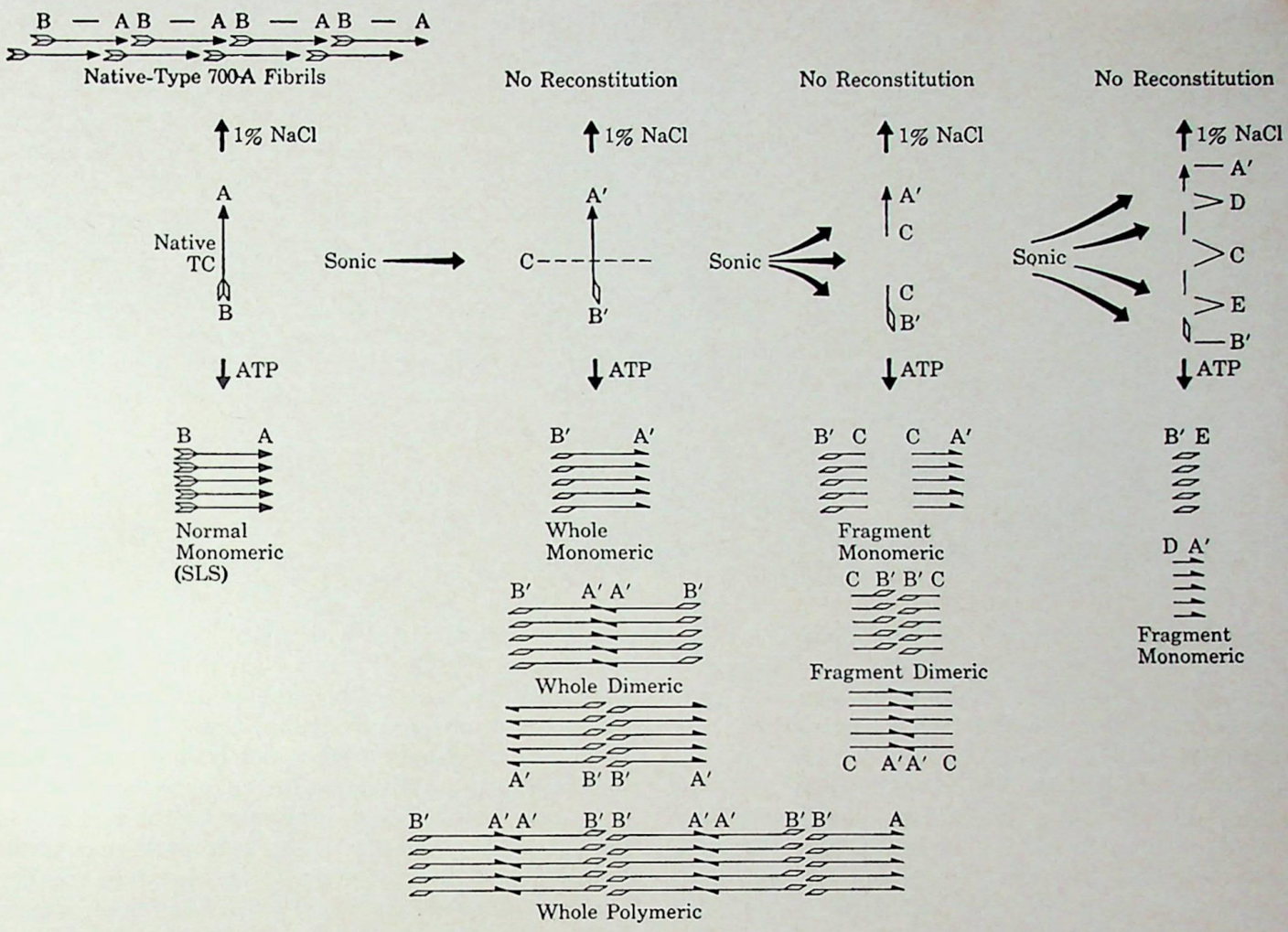


FIG. 7. Diagrammatic illustration of the chief effects of sonic irradiation on the tropocollagen macromolecules [from A. J. Hodge and F. O. Schmitt, Proc. Natl. Acad. Sci. U. S. 44, 418 (1958)].

by reference to the band patterns of the SLS-type aggregates produced by the addition of ATP to the acid solutions after irradiation.

It was discovered that sonic irradiation produces profound effects in addition to that of scission of the macromolecules into smaller fragments. The most striking of these is an alteration of "end regions," produced by relatively short periods of irradiation, without change in the length of the macromolecules. The results are shown schematically in Fig. 7, wherein the native TC macromolecules are represented by an arrow with A and B ends, indicative of the asymmetric distribution of amino acid residues reflected in the SLS type of aggregation pattern. It is thus possible to tell at a glance which is the A and B end of any particular SLS; in addition to the specific band pattern at each end, the polarization of the TC is shown at once by the position of the broad, slightly off-center interband (labeled *F-G* by Schmitt, Gross, and Highberger[23]). As was indicated earlier, the formation of the native type (700-A axial repeat) involves the formation of protofibrils that are actually linear polymers of TC by end-to-end interaction of the A-B type; lateral aggregation of such protofibrils occurs in a manner such that adjacent protofibrils are displaced axially with respect to one another by a quarter of a macromolecular length (*ca* 700 A). It is this A-B type of interaction of macromolecular ends that is first affected by sonic irradiation. In the diagrammatic representation, the altered ends are designated A′ and B′.

Following irradiation, sufficient to prevent the formation of native-type fibrils (tested for by dialysis *vs*

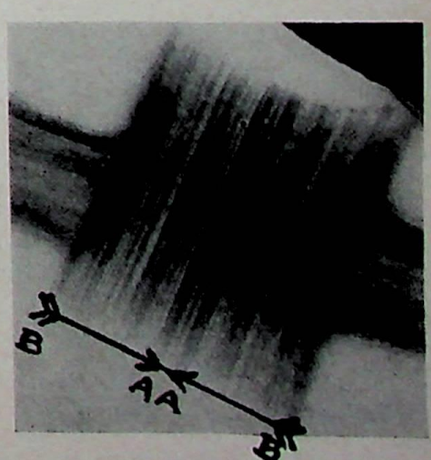


FIG. 8. Dimeric aggregation form of tropocollagen macromolecules of the type

B′—A′—A′—B′.

Produced by sonic irradiation of calfskin collagen for 20 min [from A. J. Hodge and F. O. Schmitt, Proc. Natl. Acad. Sci. U. S. 44, 418 (1958)].

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

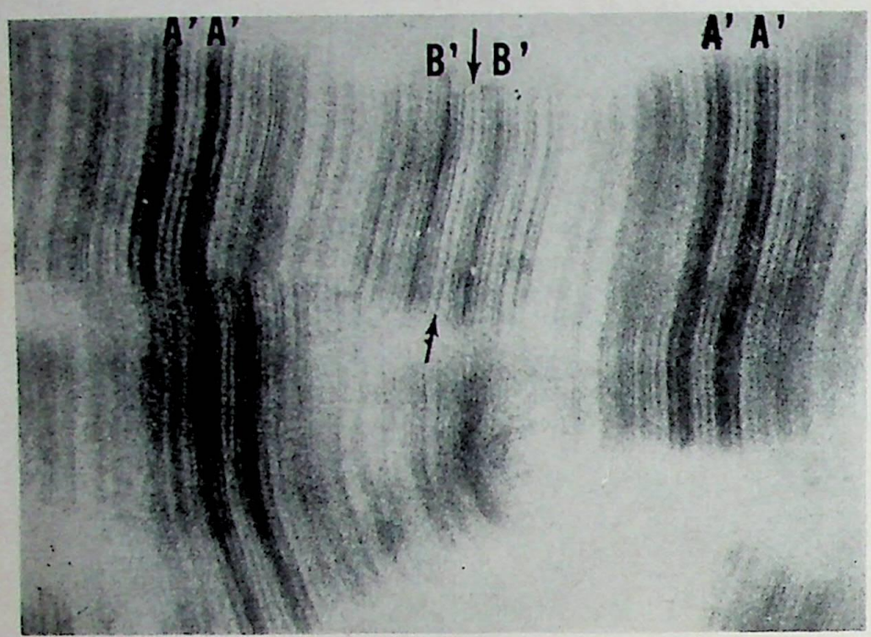


FIG. 9. Whole polymeric aggregation types of SLS aggregates from a solution of calfskin collagen treated with sonic irradiation for 240 min. Locus of A′ and B′ ends of macromolecules labeled. Arrow points to dense band at the junction between macromolecules at the B′ end [from A. J. Hodge and F. O. Schmitt, Proc. Natl. Acad. Sci. U. S. 44, 418 (1958)].

1% NaCl), i.e., by alteration of macromolecular ends, two pronounced changes are found in the ATP precipitates: (1) an increased side-to-side interaction, producing a highly exaggerated lateral aggregation into long ribbons of SLS forms; and (2) a progressive increase in the amount of A′—A′ and B′—B′ types of interaction. As a result, the formation of dimeric and polymeric forms is favored (see Figs. 8 and 9). With longer irradiation, the macromolecules are fragmented, the locus of the scission being indicated by the band pattern of the fragments.

It is noteworthy that end-to-end polymerization of scission products never involves ends produced by the fragmentation (such as those designated C, D, or E in Fig. 8). Apparently, the original ends of macromolecules are different from those produced by sonic scission. From the high density of bands (i.e., regions which combine preferentially with phosphotungstic acid) in end regions of SLS, it seems clear that certain amino-acid side chains (possibly the guanidino groups of arginine, as suggested by Kühn, Grassmann, and Hofmann,[22] or the ε-amino groups of lysine) may be concentrated in the end regions of the macromolecule.

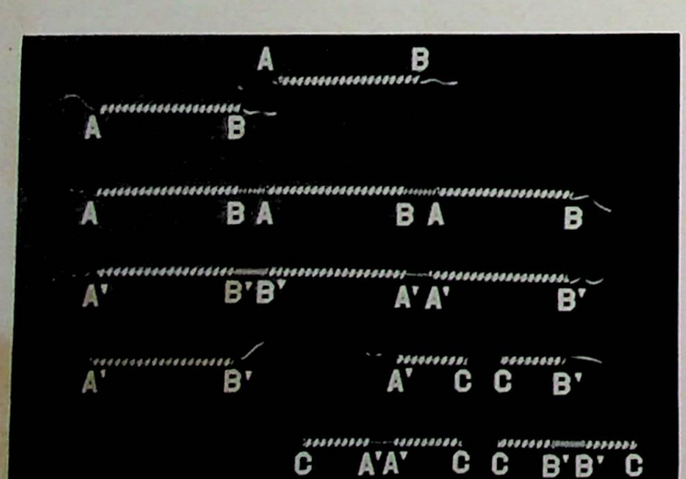


FIG. 10. Diagrammatic illustration of suggested end-to-end interaction of normal collagen macromolecules by formation of mutual coiling of terminal chains (at A-B junctions). Abnormal A′—A′ and B′—B′ junctions in sonic irradiated preparations. Scission of chains at C illustrates effect of longer periods of sonic irradiation. Note that no C—C linkage occurs [from A. J. Hodge and F. O. Schmitt, Proc. Natl. Acad. Sci. U. S. 44, 418 (1958)].

Because of the special properties of the end regions in end-to-end polymerization, it is important to obtain evidence concerning the structure in these regions. Clues are afforded by a study of the band fine structure in A′—A′ and B′—B′ linkages in the dimeric and polymeric forms. As shown in Fig. 9, the first bands at the A′ ends are separated by a region, about 100 A long, which is a typical interband (i.e., shows no dense band, hence presumably contains relatively few side chains reacting with the phosphotungstic-acid "stain"). In the case of the B′—B′ junctions, however, the separation between the first bands is about 180 A, and a darkly staining band occurs in the middle of the junctional region.

This behavior is highly suggestive concerning the nature of macromolecular ends and of end-to-end polymerization of macromolecules as follows: (1) chain appendages may occur at both ends of the native TC macromolecule; (2) these appendages may have lengths of about 100 A and 200 A at the A and B ends, respectively; (3) the amino-acid composition of the terminal chain appendages resembles typical interband regions—i.e., lacking concentrations of basic amino-acid side chains thought to characterize the band regions (except for a portion of the B end as mentioned in the foregoing).

Such considerations led Hodge and Schmitt[37] to suggest that normal end-to-end polymerization of TC monomers may involve a specific type of coiling of the terminal chains at A and B ends about each other to form a highly ordered, possibly helical, structure (see

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

Fig. 10). The "coiling energy" of such interaction in fact may be represented by the difference between the thermal shrinkage and denaturation temperatures (found by Doty and Nishihara[15] to be constant and equal to 29°C in three different types of collagen). It seems probable that such end-chain interaction involves primarily rather weak, hydrogen bonding. This would be consistent with the findings of Gross[25] that thermal gelation of salt solutions of collagen are inhibited by hydrogen-bond breakers, such as urea. In addition to such weak bonds, more stable bonds may be formed between terminal chains in the collagens of certain kinds of connective tissue, such as those which resist acid solution.

DISCUSSION

It is apparent from the behavior of TC macromolecules already described that the interaction between these macromolecules and the various types of ordered structures that result from such interaction is determined by two important factors:

(1) The specific properties and reaction potentialities that are built into the macromolecules by virtue of the linear sequence of amino-acid residues in the covalent chains, the number of chains in the macromolecule and the specific type of coiling of these chains.

(2) The chemical environment of the macromolecules which may evoke one or another of the various types of interaction patterns made possible by virtue of the internal structure of the macromolecules.

A few fibrous proteins other than collagen have been investigated along similar lines and results consistent with the foregoing conclusions obtained. Thus, Hodge (p. 409) found paramyosin to be a long (*ca* 1400 A) macromolecule capable of forming an FLS type of structure when packed in antiparallel array, though in the native fiber giving rise to a 145-A repeat or a larger period five times this long. Tropomyosin also appears to behave somewhat similarly (Hodge[38]). Also, the repeat pattern in fibrin fibrils is considerably less than the length of the fibrinogen molecules (Hawn and Porter[39] and Hall[40]), and it has been suggested that the latter are staggered with respect to molecular ends (Ferry *et al.*[41]).

Glimcher, Hodge, and Schmitt[42] presented evidence in support of the concept that the initiation of mineralization involves the nucleation of inorganic crystals by a precise juxtaposition of groups in the organic matrix to form specific stereochemical arrays. In the nucleation of hydroxyapatite in calcification, only when the purified TC macromolecules were precipitated as native-type fibrils (640-A repeat) did the system induce the formation of hydroxyapatite crystals when exposed to otherwise stable solutions of calcium and phosphate ions. If the TC macromolecules were allowed to assume other types of aggregation patterns, no nucleation occurred. In the most general case, an example is seen here of how specific macromolecular interaction serves to govern very fundamental processes, such as that of the deposition of inorganic material of specific crystalline form.

If macromolecules also possess enzymatically active sites, it is obvious that even more-complex interaction behavior may occur, particularly if products of the enzyme reaction themselves strongly influence the behavior of one or more macromolecular species in the system (e.g., ATPase in myosin).

Crucial clues to many vital biological processes, such as those briefly mentioned at the beginning of this paper, eventually may be found if such biophysical and biochemical properties of elongate macromolecules are kept in mind.

BIBLIOGRAPHY

[1] W. A. Engelhardt and M. N. Ljubimova, Nature **145**, 668 (1939).

[2] W. A. Engelhardt, Bull. Acad. Sci. USSR Ser. biol. **5**, 182 (1945).

[3] P. Weiss, Int. Rev. Cytol. **7**, 391 (1958).

[4] K. H. Gustavson, *The Chemistry and Reactivity of Collagen* (Academic Press, Inc., New York, 1956).

[5] J. H. Highberger in *The Chemistry and Technology of Leather*, F. O'Flaherty, W. T. Roddy, and R. M. Lollar, editors (Rheinhold Publishing Corporation, New York, 1956), p. 65.

[6] R. E. Tunbridge, editor, *Connective Tissue: A CIOMS Symposium* (Blackwell Scientific Publications, Oxford, England, 1957).

[7] G. Stainsby, editor, *Recent Advances in Gelatin and Glue Research* (Pergamon Press, London, 1958).

[8] F. O. Schmitt, C. E. Hall, and M. A. Jakus, J. Cellular Comp. Physiol. **20**, 1 (1942).

[9] W. T. Astbury, J. Int. Soc. Leather Trades' Chemists **24**, 69 (1940).

[10] J. T. Randall, J. Soc. Leather Trades' Chemists **38**, 362 (1954).

[11] F. H. C. Crick and A. Rich, Nature **176**, 780 (1955).

[12] A. Rich and F. H. C. Crick, Nature **174**, 915 (1955).

[13] F. H. C. Crick and A. Rich, see reference 7, p. 20.

[14] P. Gallop (personal communication, 1958).

[15] P. Doty and T. Nishihara, see reference 7, p. 92.

[16] V. N. Orekhovitch and V. O. Shpikiter, Science **127**, 1371 (1958).

[17] R. S. Bear, O. E. A. Bolduan, and T. P. Salo, J. Am. Leather Chemists' Assoc. **46**, 107 (1951).

[18] R. S. Bear, J. Am. Chem. Soc. **64**, 727 (1942).

[19] R. S. Bear, J. Am. Chem. Soc. **66**, 1297 (1944).

[20] R. S. Bear and R. S. Morgan, see reference 6, p. 321.

[21] R. S. Bear in *Advances in Protein Chemistry*, M. L. Anson, K. Bailey, and J. T. Edsall, editors (Academic Press, Inc., New York, 1952), Vol. VII, p. 69.

[22] K. Kühn, W. Grassmann, and U. Hofmann, Naturwissenschaften 44, 538 (1957).

[23] F. O. Schmitt, J. Gross, and J. H. Highberger in *Fibrous Proteins and Their Biological Significance*, Symposia Soc. Exptl. Biol. **9**, 148 (1955).

[24] F. O. Schmitt, Proc. Am. Phil. Soc. **100**, 476 (1956).

[25] J. Gross, J. Biophys. Biochem. Cytol. **2**, 261 (1956).

[26] S. G. Tomlin and C. R. Worthington, Proc. Roy. Soc. (London) **A235**, 189 (1956).

[27] H. Boedtker and P. Doty, J. Am. Chem. Soc. **78**, 4267 (1956).

[28] C. E. Hall, J. Biophys. Biochem. Cytol. **2**, 625 (1956).

[29] C. E. Hall, Proc. Natl. Acad. Sci. U. S. **42**, 801 (1956).

[30] C. E. Hall and P. Doty, J. Am. Chem. Soc. **80**, 1269 (1958).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

[31] V. N. Orekhovitch, A. A. Tustanowsky, K. D. Orekhovitch, and N. E. Plotnikova, Biokhimija **13**, 55 (1948).
[32] R. D. Harkness, A. M. Marko, H. M. Muir, and A. Neuberger, Biochem. J. **56**, 558 (1954).
[33] D. S. Jackson, see reference 6, p. 62.
[34] J. Gross, J. H. Highberger, and F. O. Schmitt, Proc. Natl. Acad. Sci. U. S. **41**, 1 (1955).
[35] J. Gross, see reference 6, p. 45.
[36] T. Nishihara and P. Doty, Proc. Natl. Acad. Sci. U. S. **44**, 411 (1958).
[37] A. J. Hodge and F. O. Schmitt, Proc, Natl. Acad. Sci. U. S. **44**, 418 (1958).
[38] A. J. Hodge. Proc. Natl. Acad. Sci. U. S. **38**, 850 (1949).
[39] C. V. Z. Hawn and K. R. Porter, J. Exptl. Med. **86**, 285 (1947).
[40] C. E. Hall, J. Biol. Chem. **179**, 857 (1949).
[41] J. D. Ferry, S. Katz, and J. Tinoco, J. Polymer Sci. **12**, 509 (1954).
[42] M. J. Glimcher, A. J. Hodge, and F. O. Schmitt, Proc. Natl. Acad. Sci. U. S. **43**, 860 (1957).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

42

Molecular Biology of Mineralized Tissues with Particular Reference to Bone*

MELVIN J. GLIMCHER†

Department of Biology, Massachusetts Institute of Technology, Cambridge 39, Massachusetts

INTRODUCTION

THERE are several reasons for presenting the topic of biological mineralization. In the first place, the organization of these tissues, from the macromolecular level to the level of the macroscopic tissue elements, offers superb examples both of structural order in biological systems and of the relation between tissue architecture and tissue function. Secondly, the *process* of mineralization illustrates the way in which the physiological function of a tissue or organ may be interpreted on the basis of the molecular structure and macromolecular organization of its components, and in terms of fairly well-characterized physicochemical phenomena. Thirdly, many of the unsolved phenomena in mineralization provide challenging problems, particularly for those with training in the physical sciences.

Certain tissues of the organism must perform a variety of mechanical functions, such as the maintenance of the form and shape of the organism against the forces of gravity, and the protection of certain delicate organs (or the organisms themselves) by enclosing them in rigid vaults. As sites of attachments for muscles and by virtue of articulations, they also provide the organism with a system of movable but structurally rigid levers. Thus, certain organisms have not only definite form and shape, but also flexibility and a means for locomotion and prehension.

Nature has devised a number of different ways of differentiating these specialized tissues. In the case of certain insects, the polymeric chitinous-protein procuticle becomes highly crosslinked providing the structural properties required for the exoskeleton. In the case of certain *Elasmobranchii*, such as the shark, a fibrous, gel-like structure, cartilage, provides both resiliency and structural rigidity. A third method, the subject of this paper, is the deposition of a substantial amount of inorganic crystals within an organic matrix: *tissue mineralization*. This process is widespread in biology (Table I), both in the plant and in the animal kingdoms, ranging from the most primitive to the most highly ordered species. Classical examples are the exoskeletons of certain marine mollusks (the shells of clams, oysters, etc.) and the endoskeletons of the vertebrates (calcified cartilage, and bone). Also, a combination of methods is used, such as the crosslinking of the chitinous-protein shell of the lobster and the collagen matrix of bone, both of which are also mineralized.

The inorganic crystals not only serve to confer new structural properties on the tissues, but also provide a storehouse of inorganic ions which may be used to help maintain the constancy of the ionic environment of the organism.

Space does not permit the discussion of the molecular biology of all of the mineralized tissues. Illustrated here are some of the major points and problems in one representative tissue, bone.

COMPOSITION AND MOLECULAR STRUCTURE OF THE MAJOR COMPONENTS OF BONE

Analytically, on a dry-weight basis, bone consists roughly of 65 to 70% of the inorganic crystals of the calcium-phosphate salt, apatite, and 30 to 35% of organic matrix of which collagen makes up the major fraction (95 to 99%).[1,2] The collagen in bone appears to be *structurally* similar to that in other tissues, although it is probably highly crosslinked.[3] The other components in the organic matrix include a number of ill-defined proteins, as well as the acid mucopolysaccharides, and constitute part of the ground substance. Some of the mucopolysaccharides probably are present as mucoproteins (noncollagenous protein complexes). The exact anatomical location of these components at a macromolecular level is not certain, and their state of aggregation and polymerization is not well known.

Although it had been known for nearly two hundred years that bone contains calcium and phosphate, it was not until 1926 that DeJong[4] demonstrated by x-ray diffraction that the crystal structure was similar to that of the apatites, more specifically hydroxyapatite [$Ca_{10}(PO_4)_6(OH)_2$], the structure of which is illustrated in Fig. 1. The exact nature of the apatite in bone, however, is still being debated.[5–11] This is partly because of the very broad x-ray diffraction reflections resulting from the extremely small crystal size, preventing crystallographic differentiation between a number of very similar proposed structures.

Further difficulties arise because the stoichiometry

* These studies were aided by research grants E-1469 (C1) from the National Institute of Allergy and Infectious Diseases, and A-2317 from the National Institute of Arthritis and Metabolic Diseases of the National Institutes of Health, Public Health Service, U. S. Department of Health, Education, and Welfare, and by research grant 12A from the Orthopedic Research and Education Foundation. Many of the major concepts involved have already been dealt with in an earlier paper [M. J. Glimcher, A. J. Hodge, and F. O. Schmitt, Proc. Natl. Acad. Sci. U. S. 43, 860 (1957)]. The detailed evidence and documentation for these concepts are currently in the press.

† Special Postdoctoral Research Fellow of the National Heart Institute, National Institutes of Health, Public Health Service, U. S. Department of Health, Education, and Welfare.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

TABLE I. Examples of biologically mineralized tissues.

Species	Tissue mineralized	Crystal chemistry	Mineral form	Major organic matrix components
Plants	Cell wall	$CaCO_3$	Calcite	Cellulose, Pectins, lignin
Radiolariens	Exoskeleton	$SrSO_4$	Celestite	(?)
Diatoms	Exoskeleton	Silica	(?)	Pectins
Mollusks	Exoskeleton	$CaCO_3$	Calcite, aragonite	Chitin, protein
Arthropods	Exoskeleton	$CaCO_3$	Calcite, aragonite	Chitin, protein
Vertebrates	Endoskeleton			
	Bone	$Ca_{10}(PO_4)_6(OH)_2$	Hydroxyapatite	Collagen
	Cartilage	$Ca_{10}(PO_4)_6(OH)_2$	Hydroxyapatite	Collagen, acid mucopolysaccharides
	Tooth			
	Dentin	$Ca_{10}(PO_4)_6(OH)_2$	Hydroxyapatite	Collagen
	Cementum	$Ca_{10}(PO_4)_6(OH)_2$	Hydroxyapatite	Collagen
	Enamel	$Ca_{10}(PO_4)_6(OH)_2$	Hydroxyapatite	Eukeratin

of the bone mineral (and many synthetically prepared apatite crystals) departs from the theoretical value of hydroxyapatite (lower Ca/P ratio). Because of their extremely small size and, therefore, large surface area, attempts have been made to explain this nonstoichiometry on the basis of the surface adsorption of excess phosphate on stoichiometric hydroxyapatite.[12,13] Aside from a number of theoretical objections to this explanation, direct experimental evidence has not verified this hypothesis.[11,14]

Other investigators have felt that lattice defects account for the aberrant stoichiometry.[6,11] In most instances, these investigators have proposed that calcium atoms were missing from the atomic structure (either internally or on the surface) and were replaced either by other cations, water, or by protons (as hydrogen bonds between the oxygens of orthophosphate groups and protons). The latter have been referred to as defect apatites.[11] However, defect crystals in which there are vacancies in certain lattice positions are ones in which the order of magnitude of these "holes" or "defects" is possibly one per thousand or ten-thousand atoms. To satisfy the requirements in this particular case, however, one or even two calcium atoms out of every ten would have to be absent from the lattice structure. It is very doubtful that so many calcium atoms could be absent or substituted for without some structural change in the lattice.

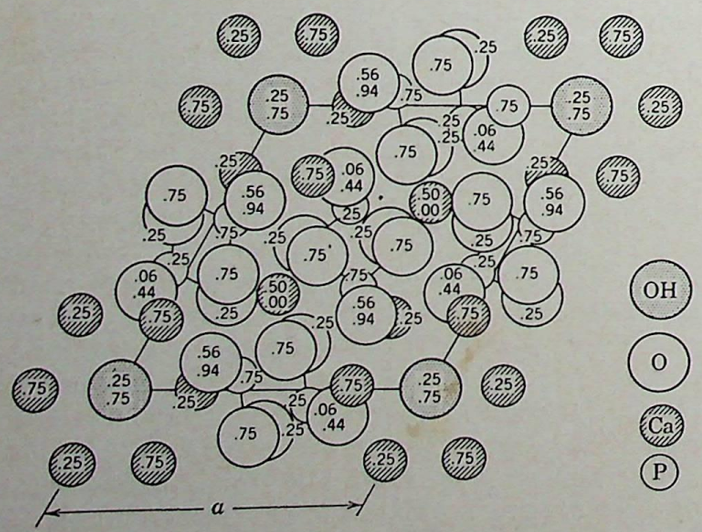


FIG. 1. Atomic arrangement of the constituents of hydroxyapatite projected on the 001 plane. The numbers refer to the fractional height in the unit cell of the atoms in the plane perpendicular to the paper (*c*-axis), as reported by Carlström[6] [from A. S. Posner, Clin. orthoped. **9**, 5 (1957)].

Such structural changes have been postulated in the case where water or hydronium ions replace Ca atoms in the lattice structure. Experiments have shown (1) that a considerable amount of water is lost both from bone apatite crystals and *in vitro* prepared apatite crystals when subjected to progressively increasing temperatures after initial dehydration,[8–10,15] and (2) that apatite crystals precipitated from aqueous solutions have associated with them an amount of water many times greater than that adsorbed by initially dry crystals from a vapor phase of water.[16] This "excess" water cannot be separated from the crystals by mechanical centrifugation at 80 000 g.[16]

Since the structure of hydroxyapatite allows for no water of crystallization, several explanations have been offered to account for this "excess" water.

Thus, the substitution of water (or hydronium ions) for Ca atoms, would not only explain the low Ca/P ratios, but would also be consistent with the data on water discussed above.

The objection that such a relatively large replacement of Ca atoms would lead to structural changes in the lattice (*vide supra*) is acknowledged by one group which believes that there *are* structural differences between hydroxyapatite and the bone apatite crystals, and refers to the latter as α-tricalcium phosphate or hydrated tricalcium phosphate without specifying either the exact differences in the structural parameters or the position of the water in the new structure.[8–10,17]

Considering the water as specifically replacing columnar calcium atoms in the apatite lattice, it has been proposed also that certain low-ratio apatites are layered structures consisting of "sheets" of hydroxyapatite held together by hydrogen bonds between the phosphate groups and water.[18]

This layered apatite structure referred to as octocalcium phosphate (OCP) gave an almost identical

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

x-ray diffraction pattern as that of "pure" hydroxyapatite except for several additional reflections,[5,17] presumably thought to be owing to the superlattice effect of the layered water.

On the other hand, the "excess" water (particularly in the case of *in vitro* precipitated apatite crystals in aqueous solutions) has also been postulated to be the result of a large hydration shell composed of about 60 molecular layers of water "bound" to the crystals because of an electrical-field asymmetry of the crystal surfaces.[16] This amount of water is many times greater than that adsorbed by initially dry crystals from a vapor phase of water and consists of many more molecular layers of water than usually considered possible for surfaces to bind as a result of electrical-field effects.

It appears possible from the data on the replacement of calcium atoms by water (or hydronium ions) that, when apatite crystals are precipitated in aqueous solution, part of the "excess" water may be accounted for by its incorporation *into the crystal lattice*, resulting in a phase change of the solid.

Another factor in determining the amount of water associated with apatite crystals, particularly those precipitated in aqueous solutions, is the size, shape, and interaction properties of the crystals which are discussed later in this paper (page 386). Recently, however, single crystals of OCP have been prepared, and it has been shown to be a distinct compound whose structure, while closely related, is not identical to hydroxyapatite.[19] These investigators also felt, however, that OCP was also a layered structure with the layers separated by water molecules.

Dehydration of OCP caused a characteristic x-ray reflection (18.4 A) to shift to progressively lower spacings and to disappear completely upon elimination of approximately two-thirds of the water. The crystals then showed a typical apatite pattern, indicating that, with loss of a certain amount of water, the OCP structure is unstable and a true phase change occurs from OCP to apatite.

Octocalcium-phosphate (OCP) crystals prepared in the laboratories of A. S. Posner, American Dental Association,[20] Research Division, National Bureau of Standards, have been well characterized by x-ray diffraction and have shown minor differences both in the lattice spacings and intensities of the reflections as compared to those of hydroxyapatite. These OCP crystals have been examined in this laboratory by electron-microscopy and electron diffraction. The electron-diffraction patterns were indistinguishable from those of hydroxyapatite indicating that, under the conditions of vacuum and temperature in the electron microscope, a phase change had occurred from an OCP lattice to an apatite lattice by the dehydration process, confirming the report[19] referred to earlier.

Another problem is the determination of the position and state of aggregation of carbonate, which has been postulated to exist in the crystal lattice *per se* (as a carbonate apatite) as a separate phase (calcium and magnesium carbonate) or adsorbed on the crystal surfaces as carbonate ions.[7–10,21,22]

These examples illustrate that, although the evidence is quite substantial that the lattice structure of the inorganic crystals in bone is either that of hydroxyapatite or of something similar, the intimate details such as the relationship between OCP, other hydrated calcium phosphate compounds, defect apatites, hydroxyapatite, and the structure and nonstoichiometry of the bone crystals and synthetically prepared "apatite" crystals is obviously not yet clear.

ANATOMICAL STRUCTURE

Figure 2(a) demonstrates the gross appearance of the upper end of a femur. The honeycomb-like appearance of the head and metaphysis is referred to as the *spongiosa* and is composed of delicate spicules of bone called *trabeculae*. The bone in the cortex of the shaft is much more densely packed and is referred to as *compact bone*. In both cases, the structure of the adult bone consists of a series of layers or lamellae. In the cortex of many long bones, the lamellae are further arranged concentrically around a central canal forming hollow "cylinders" containing small blood vessels. These small cylinders of bone are the basic units of such compact bone and are called *osteones* or *Haversian Systems* (Fig. 3). They are longitudinally oriented in the general direction of the long axes of the long bones (Fig. 4).

The collagen fibers in a lamella are arranged in small bundles which encircle the canal in continuous spirals crossing one another and resulting in a trellis-like arrangement.[23,24] Although the general direction of the bundles in the same lamella is similar, it varies from one lamella to the next, giving a characteristic appearance when viewed in polarized light.[25] This arrangement of the fibers within any one lamella and in consecutive lamellae imparts maximum structural properties to the tissue.

The inorganic crystals of apatite are deposited in this highly organized and ordered matrix of collagen and ground substance. Their distribution can be visualized by the use of microradiography (Fig. 5). Note the inhomogeneity, not only with respect to the entire section, but even within any one Haversian system.

The trabeculae, or "bone girders" of the spongiosa are oriented functionally in that they closely parallel the trajectories of maximum stress [compare Figs. 2(a) and 2(b)]. This allows the structure to resist mechanical stress and strain in the most efficient manner and with the greatest economy of material in accordance with sound engineering principles.

The *size*, *number*, *disposition*, and *orientation* of trabeculae of the spongiosa also can *change* in response to altered mechanical demands. This functional adaptation of bone to mechanical stress was emphasized first by Wolff in 1892 (Wolff's law),[26] yet, to date, no adequate explanation of the mechanism has been offered.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

The structural order evident in the arrangement of the macroscopic trabeculae and in the microscopic Haversian systems of bone is also evident at a lower order of magnitude. In Fig. 6, an electron micrograph of a longitudinal section of cortical bone, the collagen fibrils are well oriented and the 640-A axial repeat of the fibril is accentuated in many areas by the specific location of the dense inorganic crystals. The crystals are well oriented with their long axes parallel to the collagen fibrils and intimately related to them. Electron diffraction (Fig. 7) confirms that it is the crystallographic *c*-axis of the inorganic crystals which is parallel to the fibril axis of the collagen.

Most workers have felt that the inorganic crystals of adult bone were in the ground substance between the collagen fibrils,[27,28] although more recent work indicates that some of the crystals might be within the fibrils.[29] Interpretation of electron micrographs has been difficult because of the dense packing of the collagen fibrils (Fig. 5), making it impossible to say with certainty whether the crystals were within or on the surface of the collagen fibrils, or just generally associated and oriented with them by virtue of the over-all organization, orientation, and dense packing of the fibrils in the tissue. Our laboratory has recently resolved this issue by studying bone in which the collagen fibrils are widely separated, and it is possible to see the relationship between individual fibrils and individual crystals by high-resolution electron microscopy.

Figure 8, an electron micrograph at low magnification, indicates the regular arrangement of the collagen fibrils in alternate layers and the relative looseness in the packing of the individual fibrils. It is quite obvious that the crystals are *within the fibrils* and not in the intervening ground substance. Figures 9–12 are longitudinal and cross sections at higher magnifications, and more clearly emphasize the position of the inorganic crystals *within the collagen fibrils*. That this is not an initially random crystal precipitation between the fibrils, later reorganized within the fibrils by recrystallization, is evident from an examination of electron micrographs of the earliest stages of calcification in

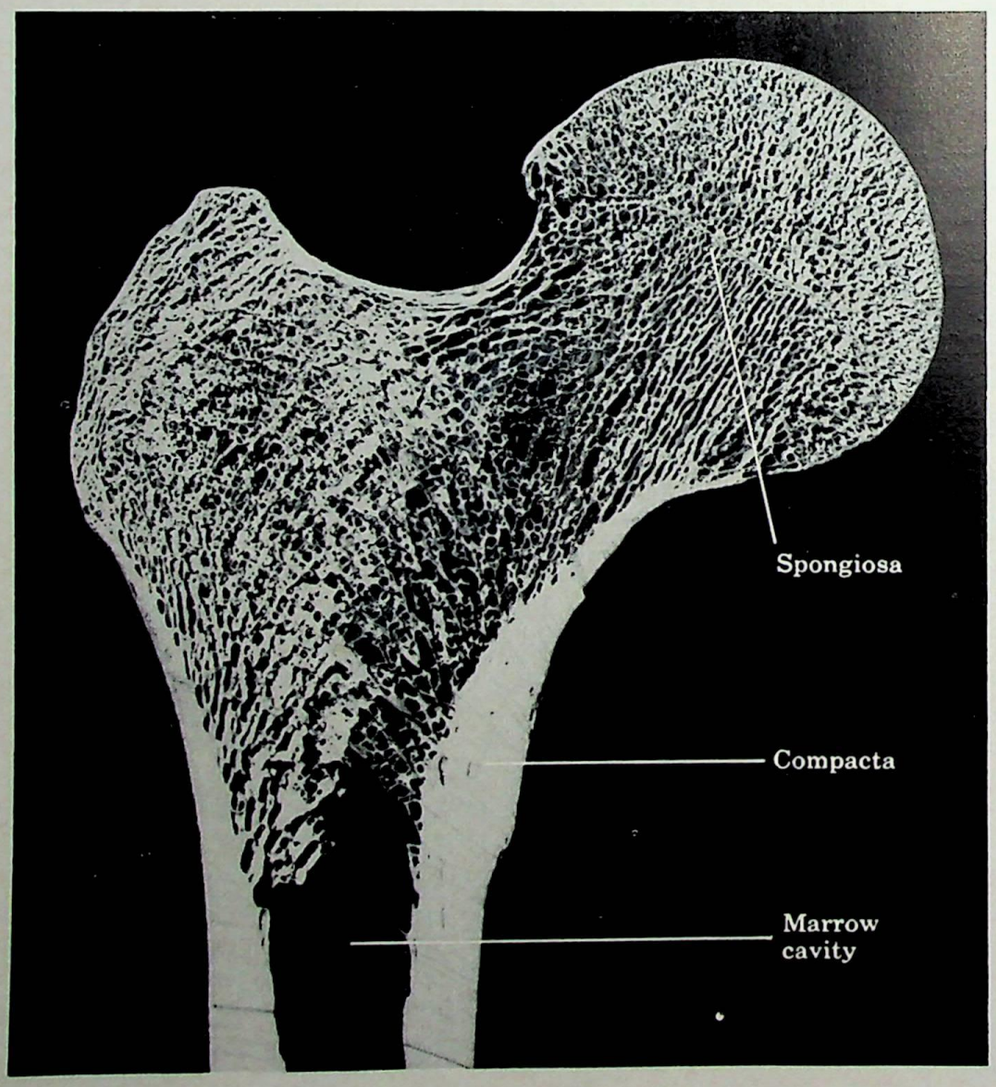


FIG. 2(a). Coronal section of the upper end of a human femur [from D. W. Fawcett, in *Histology*, R. Greep, editor (The Blackiston Company, New York, 1954), p. 133; reproduced from original kindly supplied by the author].

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

embryonic bone.[30,31] Figures 13 and 14 are electron micrographs of the initial stage of calcification in embryonic bone. These clearly show that the earliest crystals are regularly spaced and deposited *within* the collagen fibrils.

The location, distribution, and orientation of the crystals, by virtue of their position within the fibrils, provide the most efficient arrangement for effectively resisting mechanical stresses.

The extremely small size of the crystals (200 to 400 A $\times$15 to 30 A) and the fact that one and probably two of their dimensions consist of only a few unit cells, result not only in a tremendous surface area, but also in a large percentage of the atoms being in surface or near-surface positions. This easy accessibility to the crystal interior allows for a rapid exchange of ions between the crystals and the interstitial fluids.

In summary, starting with the major components themselves (collagen and apatite) which are ordered materials, the organization of the tissue, from a macromolecular level to the arrangement of the macroscopic trabeculae, constitutes a highly ordered, well-organized structure, superbly adapted both mechanically and chemically to perform its biological functions.

MECHANISM OF CALCIFICATION

Introduction

The many theories of calcification may be classified, in general, into two groups. The first, referred to as the "booster" theory,[16] proposed that a specific enzyme (or enzymes) exists in the areas undergoing calcification and that it splits off inorganic phosphate from an organic substrate.[32] This local "boosting" of the concentration of phosphate then exceeds the level of spontaneous precipitation and results in crystallization.

A second concept, first proposed in 1921[33] and recently revived, suggests that the *organic matrix* of calcifiable tissues initiates the formation of crystals. Recent biochemical evidence[34] and the inadequacy of the "booster" theory in explaining the precise localization of the

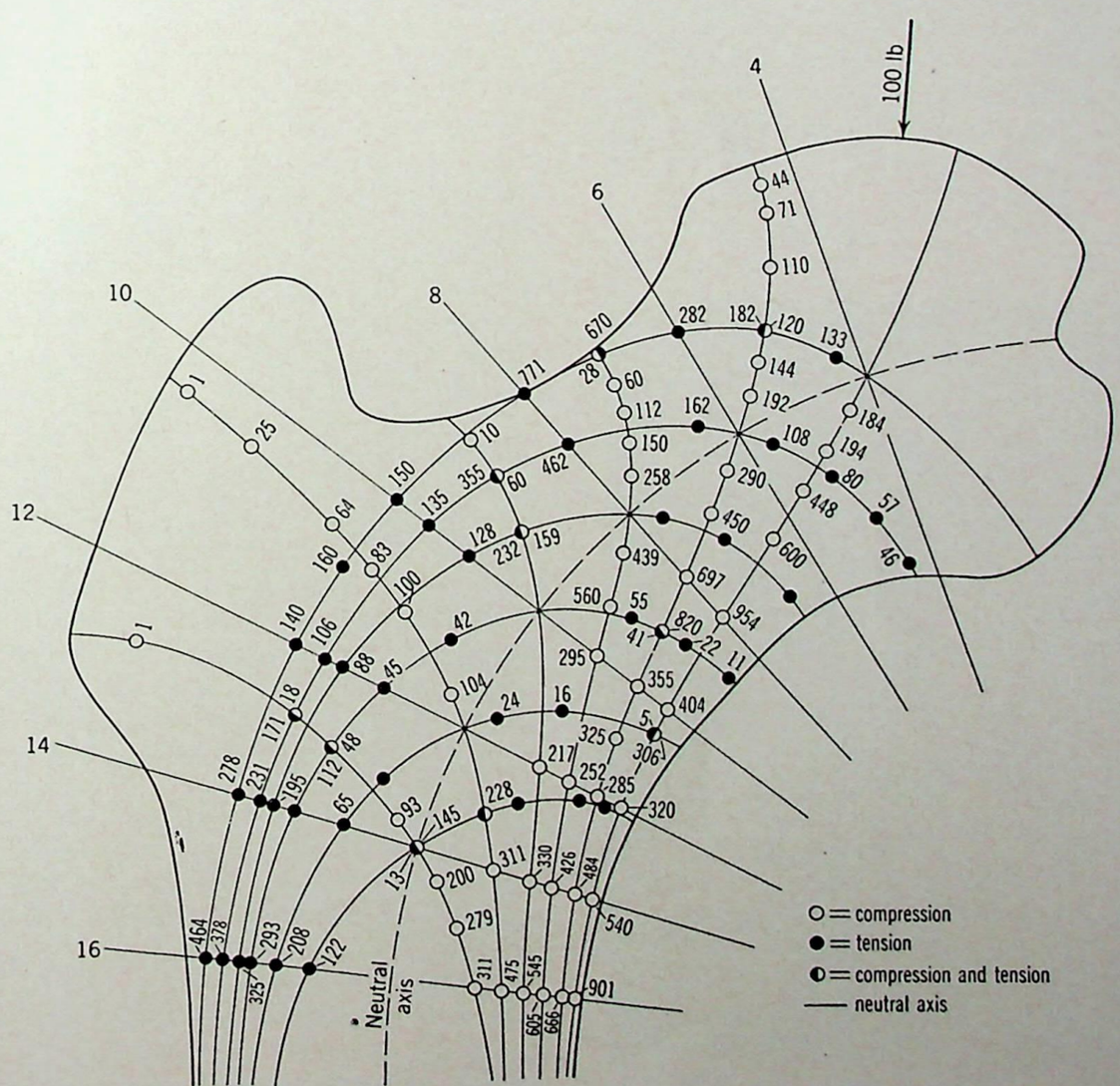


FIG. 2(b). Diagram of the lines of stress in the upper femur (after Koch) [from *Gray's Anatomy*, W. H. Lewis, editor (Lea and Febiger, Philadelphia, 1942), twenty-fourth edition, Fig. 245].

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

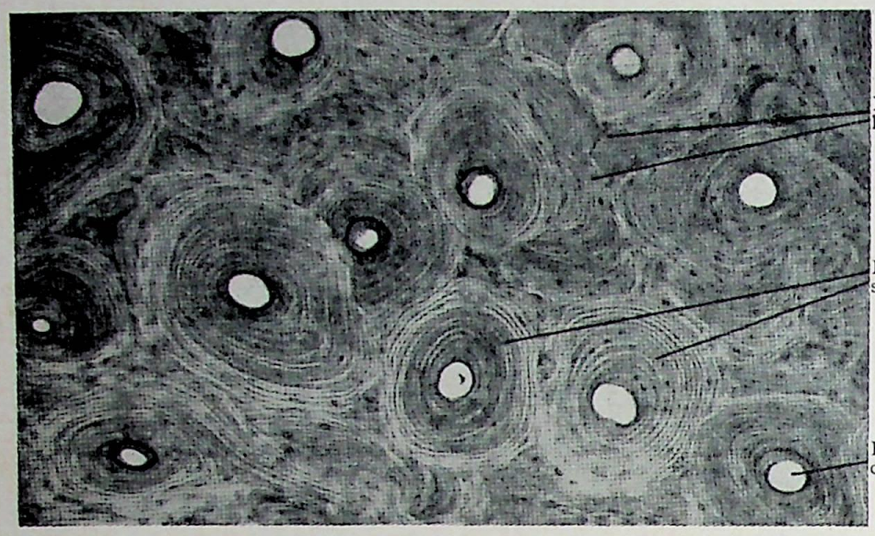


FIG. 3. Cross section of decalcified compact bone ×80. Note the alternating dark and light layers in the Haversian systems and interstitial lamellae attributable to consecutive changes in collagen-fiber orientation [from D. W. Fawcett in *Histology*, R. Greep, editor (The Blackiston Company, New York, 1954), p. 134; reproduced from original kindly supplied by the author].

crystals at an ultrastructural level strongly support the role of the organic matrix in calcification. Opinion as to what substance (or substances) in the organic matrix is responsible for the induction of crystallization and the nature of the *mechanism* of induction has varied.[35–42]

Because of the complexities in biological mineralization, it is important that the physicochemical nature of the mineralization phenomenon be clearly defined before proceeding to the experimental evidence and hypothesis which follow. Crystallization, or the formation of inorganic crystals from solutions where none previously existed, represents a *phase change*. This physical change in state can be divided arbitrarily into *crystal nucleation*, the process of forming the initial fragments of the new phase, and *crystal growth*, the subsequent growth of these fragments into clearly defined crystals. In addition, the phenomena of recrystallization (the growth of large crystals at the expense of smaller ones), may also play a role even after the solid state has been achieved. Failure to distinguish these interrelated but separate phenomena and failure to differentiate between the many *regulatory* processes controlling them and their underlying *mechanisms*, can be confusing. A brief review of some of the thermodynamic and kinetic principles related to *phase transition* is given in the Appendix at the end of this paper.

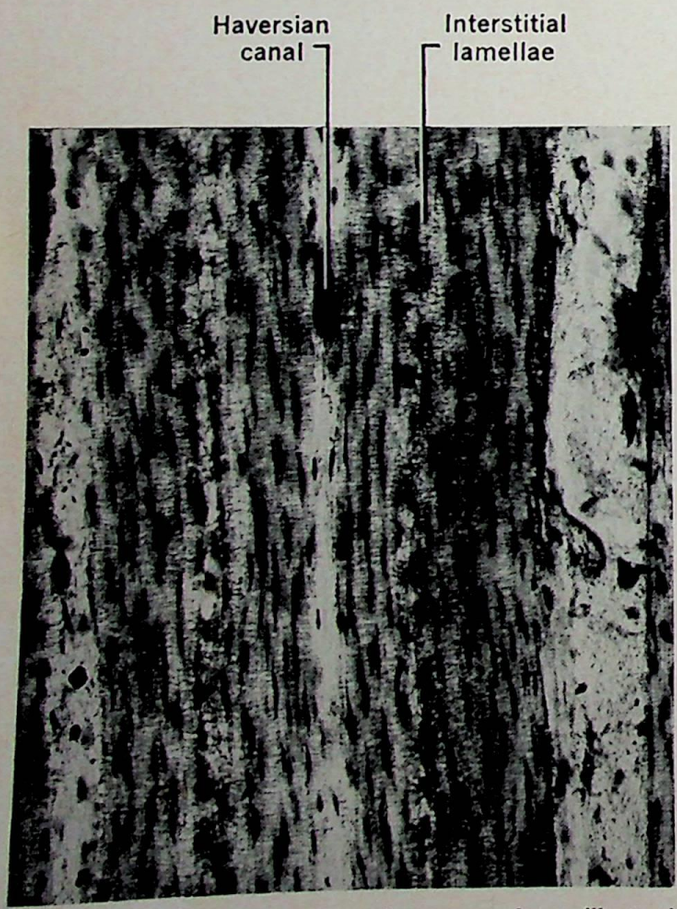


FIG. 4. Longitudinal section of compact bone illustrating Haversian systems (original magnification ×120) [from J. P. Weinmann and H. Sicher, *Bone and Bones, Fundamentals of Bone Biology* (C. V. Mosby Company, St. Louis, 1947); reproduced from original kindly supplied by the authors].

Certain misconceptions concerning the formulation of apatite crystals from solution are clarified by the phenomenological description in the Appendix regarding the manner in which phase changes occur.

It has been assumed[6,16,43,44] that, in the precipitation of apatite crystals, the formation of brushite ($CaHPO_4 \cdot 2H_2O$) must occur first, later to be hydrolyzed or otherwise converted to apatite [$Ca_{10}(PO_4 \cdot 6(OH_2)$]. This assumption has been based on the incorrect premise that an 18-body collision is required for hydroxyapatite formation whereas only a two-body collision is required for brushite formation.

Several conceptual errors need clarification. In the first place, when calcium and phosphate ions in solution aggregate to form a crystal, the change in state which occurs is a *physical change in state*—i.e., a change in the state of aggregation of the ions from a solution phase to a *solid* phase. The empirical formulas of the solids in such cases [$CaHPO_4 \cdot 2H_2O$ *or* $Ca_{10}(PO_4) \cdot 6(OH_2)$] do not refer to a molecule or molecules of brushite (containing 2 ions and 2 water molecules) or of apatite (containing 18 ions), but merely represent the *ratios* of all of the constituent ions of the *solid phase* in terms of smallest whole numbers. The ions are part of a crystal-

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

line solid containing *thousands* or usually *millions* of ions arranged in a definite spatial configuration characteristic for the particular crystal.

Secondly, even in chemical reactions where a physical change of state does not occur, neither the reaction *order* nor the reaction *mechanism* can be deduced simply on the basis of the chemical formulas of the reactants or the reaction products. In the nucleation of new phases, nuclei probably arise by the stepwise addition of single molecules, atoms, or ions[45,46], and there is no obvious association between the empirical formula of the solid and the size and composition of the nucleus or the order of the reaction.

Inspection of the empirical formulas of various possible solids in no way permits one to predict the *probability* that the formation of *one* particular solid is *more likely* than another. This, of course, will depend on the relative amount of work which is necessary for the formation of a cluster of a particular composition, size, shape, and configuration under the specific conditions of the experiment, i.e., *p*H, temperature, etc.

Some simple examples illustrate the foregoing points. The formation of ice from water may be written H_2O(liquid)$\rightarrow$$H_2O$(solid). It is obvious that one is not dealing with single "molecules" of ice, but with a *solid* containing many water molecules in a specific steric configuration. It is also apparent that neither the *size* of the critical cluster nor the *reaction kinetics* can be deduced, either from the chemical formula or from the equation which describes the change in state. Physicochemical evidence[46] suggests that the nucleus which initiates this phase change is composed of about 80 to 100 water molecules. These nuclei do not form by the simultaneous collision of 80 to 100 water molecules, but are believed to arise as the result of the stepwise addition of single water molecules. The reaction is, therefore, bimolecular in *mechanism*, and of the 80th to 100th *order*.

As a second example, the formation of sodium chloride crystals from solution may be written as Na+ (ion aqueous) +Cl− (ion aqueous)$\rightarrow$NaCl (solid). This is a physical change in state—i.e., a change in the state of aggregation of the sodium and chloride ions from ions in solution to ion constituents of a crystal. In this solid, there are no NaCl "molecules," but an array of Na+ and Cl− ions, with each sodium ion equally shared by six chloride ions and each chloride ion by six sodium

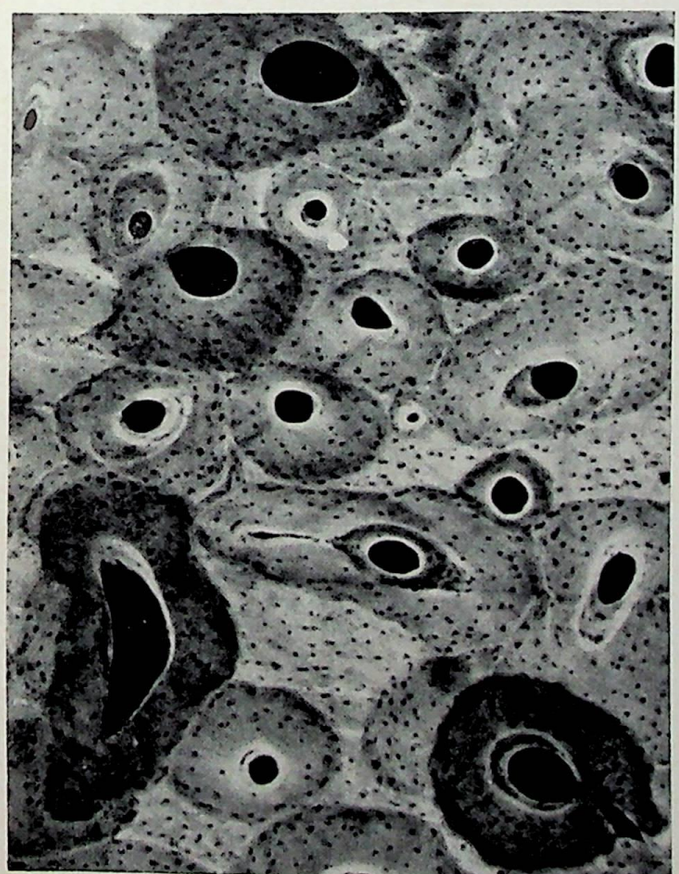


FIG. 5. A microradiogram of a thin cross section of normal cortical bone. Note the variation in density (indicating inorganic-crystal concentration) within any one Haversian system, as well as within the section as a whole. Reproduced from original kindly supplied by A. Engström, Karolinska Institutet, Stockholm, Sweden.

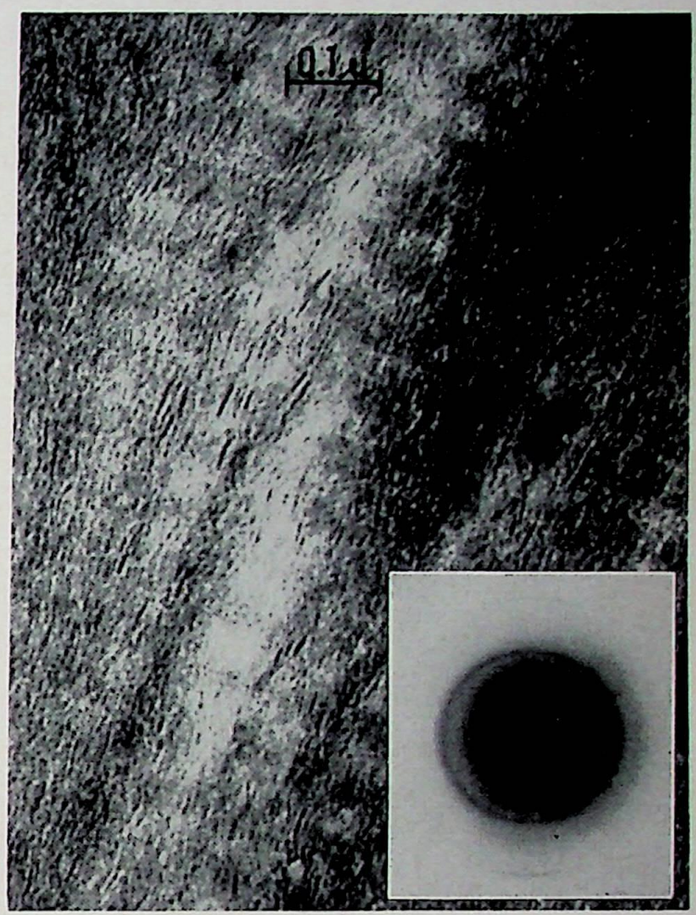


FIG. 6. Electron micrograph of a longitudinal section of compact bone. Note the accentuation of the characteristic axial repeat of the collagen fibrils in some areas by the small inorganic crystals, whose long axes approximately parallel the collagen-fiber axis. ×100 000.

FIG. 7 (lower right-hand corner). Selected-area electron diffraction of the specimen in Fig. 6, showing characteristic pattern of apatite. Arcing of the 002 and 004 reflections indicates that the crystallographic *c*-axes of the crystals are oriented parallel to the collagen-fiber axis, and, therefore, correspond to the long axes of the crystals.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

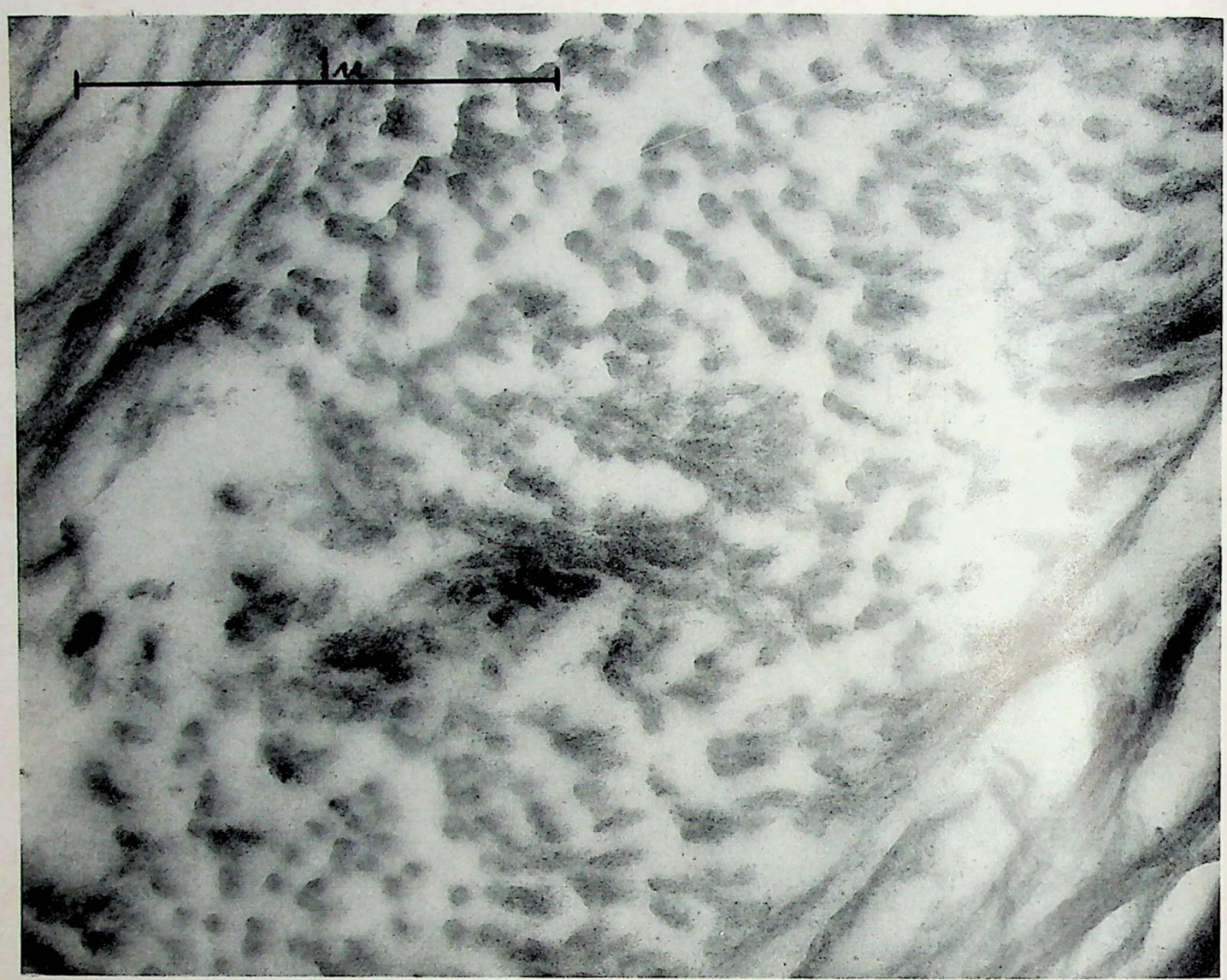


Fig. 8. Electron micrograph of a cross section of fish bone. Note the alternating direction of the collagen fibers in consecutive layers and the packing of the fibrils. Even at this low magnification it is obvious that the dense inorganic crystals are *within* the fibrils and not in the intervening ground substance. ×58 000.

ions. The chemical formula of the solid (NaCl) obviously does not indicate that this change occurs as a result of a 2-body collision.

There appears to be no doubt that, under *certain* physicochemical conditions (low *p*H, 6.0–6.2), one *can* precipitate the calcium phosphate solid, brushite, and that this can in turn be hydrolyzed or otherwise converted to apatite under the proper physicochemical conditions (raising the *p*H, for example). However, one cannot infer that, when a calcium-phosphate salt is precipitated, the formation of brushite is more probable or more likely than apatite, simply because its *empirical* formula contains fewer atoms and ions.

MECHANISM OF CRYSTAL INDUCTION

The paper by F. O. Schmitt (p. 349) notes that not only is the collagen macromolecule a "crystalline" protein, but also the collagen *fibril*—by virtue of its *ordered aggregation* of such macromolecules, and high degree of structural regularity, as seen both by electron microscopy and by low-angle x-ray diffraction—may be considered "crystalline."

These facts and the obvious association between the apatite crystals and the collagen fibrils in normally calcified tissues recommended investigation of the possibility that this intimate anatomical relationship was evidence that either the macromolecular *structure* of collagen or the macromolecular *aggregation state* of the collagen fibrils was responsible for the *induction* of calcification,[40] by acting as a catalytic heterogeneity for the nucleation of apatite crystals. Other studies had also suggested a role for collagen[38,39] or collagen chondroitin sulphate complexes in calcification.[35,36]

For such a mechanism to be operative, however, the solutions from which the crystals are formed must be in metastable equilibrium. It is, therefore, important to establish the state of equilibrium of the fluids immediately surrounding the fibrils in relation to the formation of this new phase (apatite crystals) *in vivo*. Although it has been demonstrated that the bulk of the extra-

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

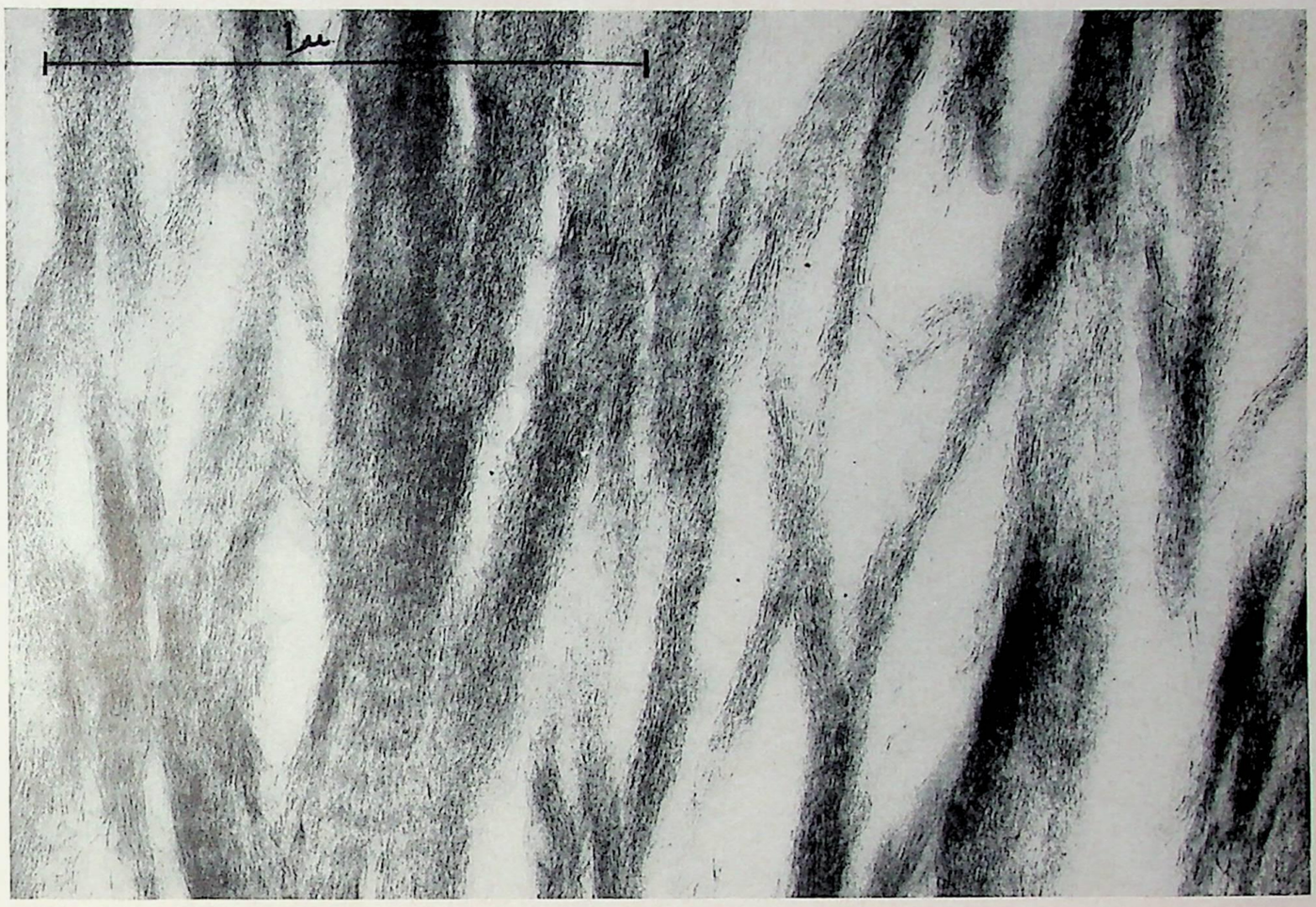


FIG. 9. Electron micrograph of fish bone in a region of loosely packed fibrils such as in Fig. 8. Despite variations in individual fibril direction, long axes of inorganic crystals are parallel to the individual collagen fibrils with which they are associated by virtue of their position within the fibrils. ×75 000.

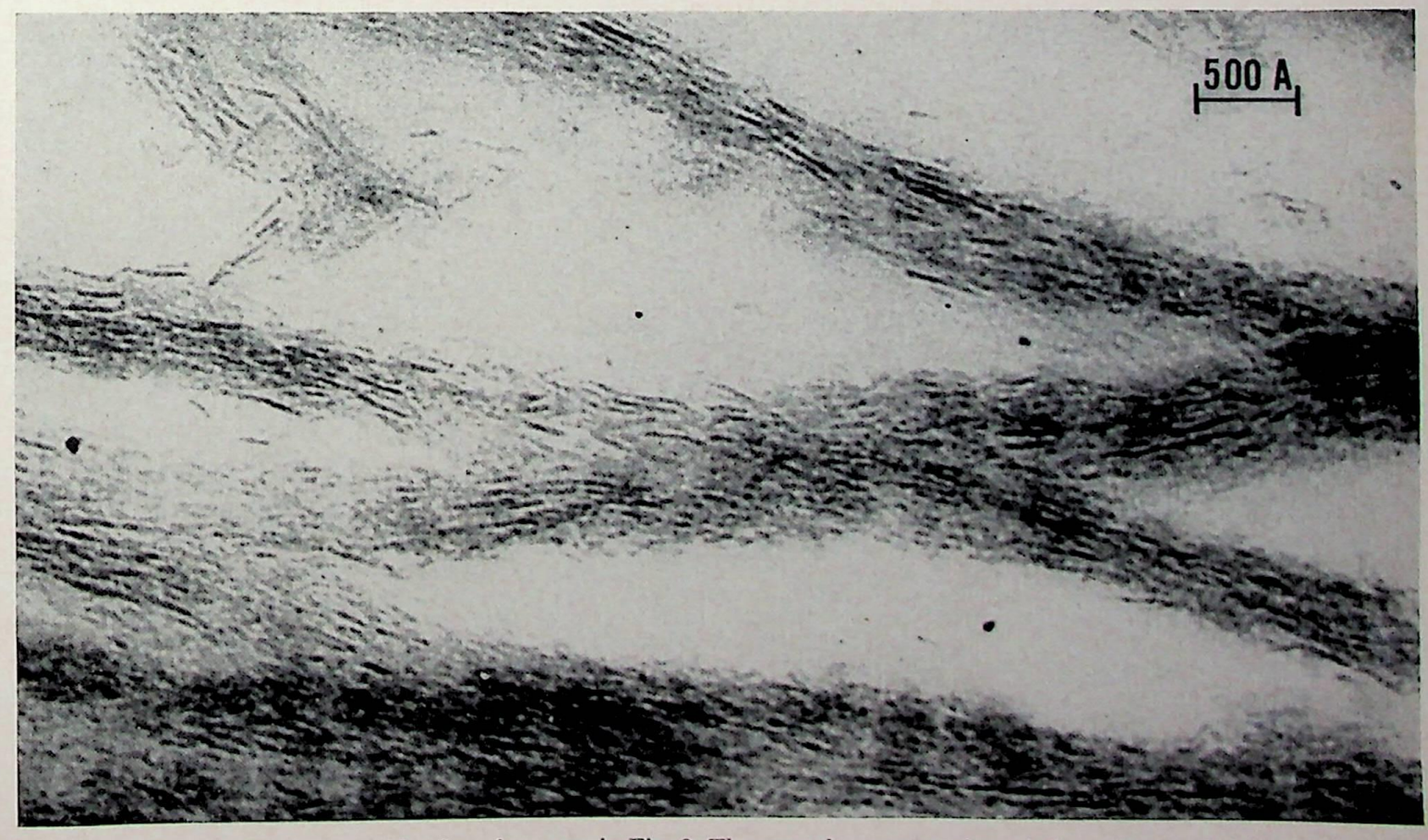


FIG. 10. Higher magnification of an area in Fig. 9. The crystals appear to be rod-shaped, approximately 200 to 400 A in length and 15 to 40 A in width. ×225 000.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by eGangotri Foundation USA

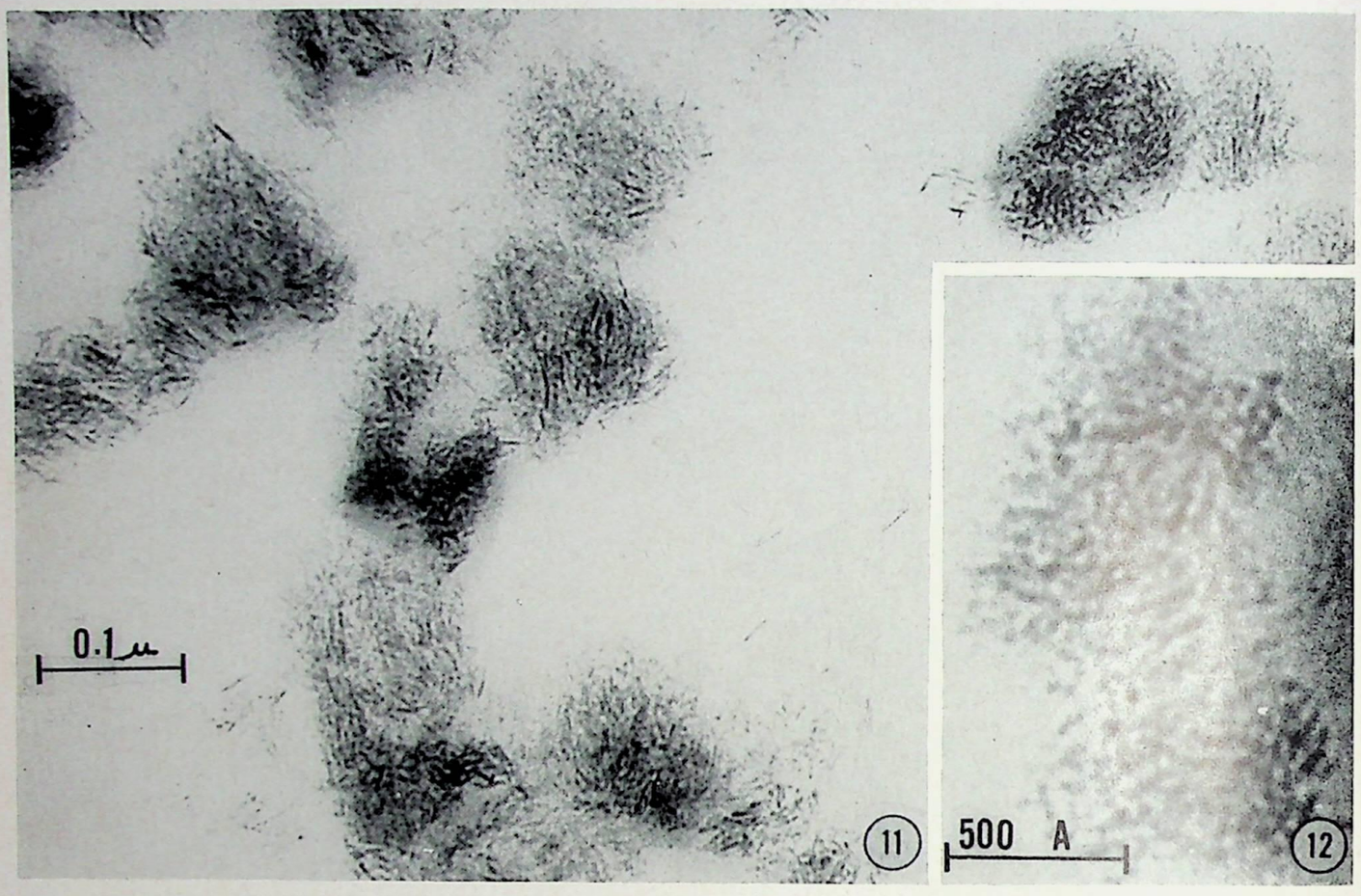


FIG. 11. Higher magnification of cross sections of fibrils in fish bone. The apatite crystals are obviously within the collagen fibrils. ×165 000.
FIG. 12. Cross section of two fibrils in an area similar to Fig. 11. The crystals have appearance of rods viewed on end and appear to be hexagonally packed. ×430 000.

cellular interstitial fluid is metastable with respect to the formation of apatite crystals,‡ the *local* values immediately surrounding the collagen fibrils in any particular tissue might be markedly different because of active cellular-controlled compositional changes, or active or passive transport and diffusion phenomena.

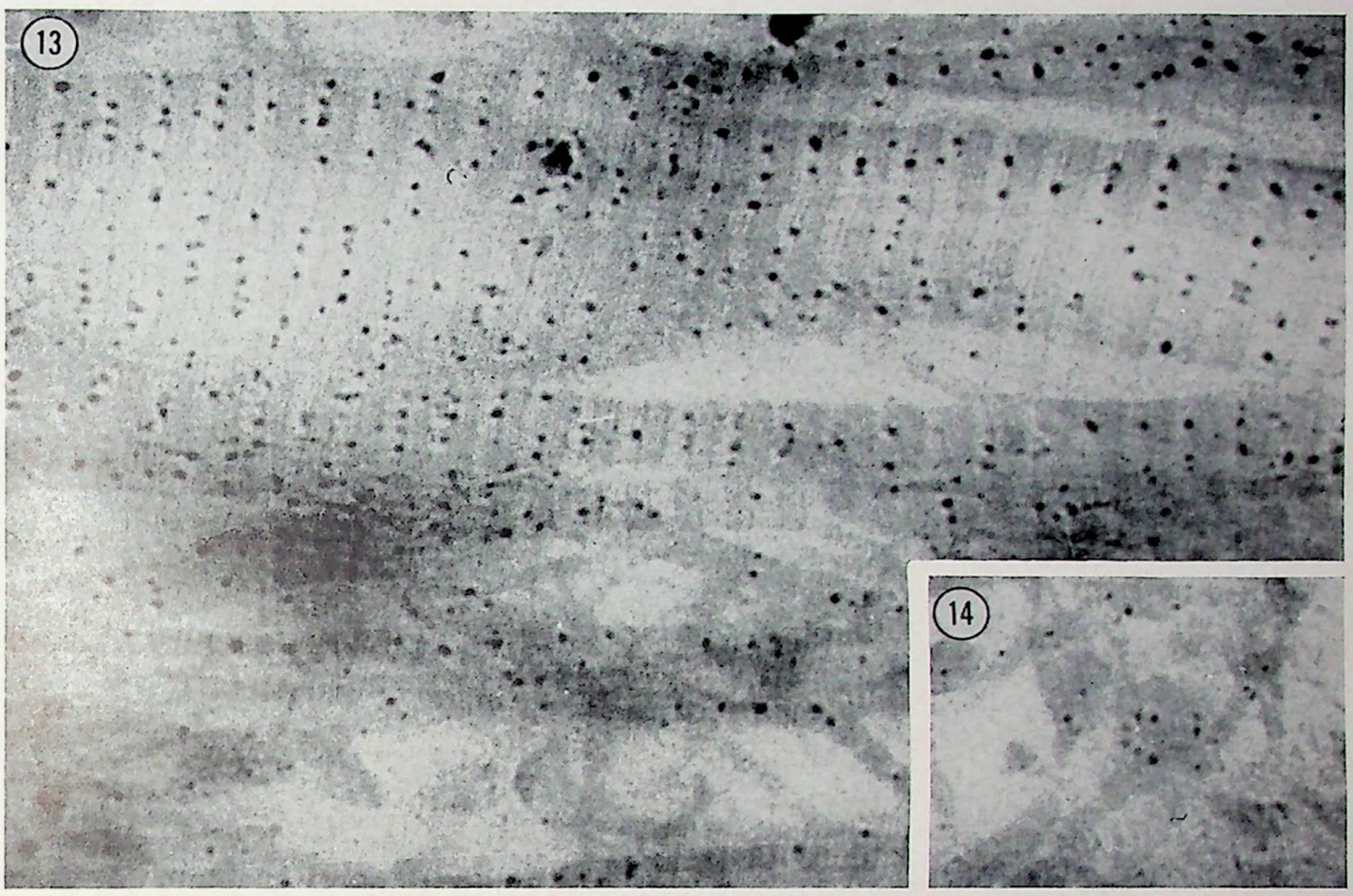
Although it is informative to examine the data with regard to the state of equilibrium of the unaltered extracelluar fluids, not only with reference to the mechanism of induction but also particularly with regard to the *regulation* and *control* of this process in other non-mineralized tissues, the *immediate objective*, however, was to establish whether or not collagen *could* induce the formation of apatite crystals from solutions which *were* metastable, and to determine the nature of this specificity (if any) and its mechanisms. The experiments discussed in the following were conducted at the Massachusetts Institute of Technology in collaboration with A. J. Hodge and F. O. Schmitt, and were designed to answer these questions.§

The collagen of many connective tissues can be dissolved in a number of weak acids and neutral buffers yielding viscous solutions of the macromolecules (see F. O. Schmitt, p. 349). These can subsequently be reaggregated and reconstituted into fibrils with the typical and characteristic axial repeat and intraperiod fine structure of native fibrils. By appropriate treatment of the connective tissue prior to dissolving the collagen and by several recrystallizations, the collagen fibrils can be prepared relatively pure, with only minute traces of ground-substance constituents.

‡ The ion product ($a_{Ca}++\times a_{HPO_4}-$) in the serum of many vertebrates is considerably higher than the same product in solutions of equivalent ionic strength, *p*H, temperature, etc., where equilibrium has been approached either through precipitation or through dissolution of apatite crystals.[38,39] Since this product appeared to be critical in determining whether or not precipitation of a solid phase occurred, it was concluded that serum and interstitial fluid were supersaturated with respect to the bone mineral. While these thermodynamic data are not conclusive, direct experimental evidence supports this supposition. Inorganic solutions with ion products ($a_{Ca}++\times a_{HPO_4}-$) similar to serum were stable for indefinitely long periods of time, but showed rapid separation of *more* solid when exposed to bone mineral[47] or to synthetic apatite crystals.[48] It would appear, therefore, that the unaltered extracellular interstitial fluid is metastable with respect to the formation of apatite crystals.

§ Although we are fully aware of the limitations of a model system which is admittedly a great deal simpler than the process as it occurs *in vivo*, it has enabled us to characterize and distinguish what appears to be the *mechanism* of crystal induction as well as several closely allied phenomena in the process of calcification and to gain an insight into their mechanisms. Many of the minute details and a quantitative description of these phenomena and the manner in which they are biologically *regulated* will, of course, have to be gathered from the intact organism. However, one must remember that it is difficult even to know *which* details to search for if the basic underlying mechanisms are not at least clearly defined conceptually.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA



FIGS. 13 and 14. Electron micrographs of avian embryonic bone in the earliest stages of calcification. Note the regular and periodic arrangement of the dense apatite crystals within the collagen fibrils [from S. Fitton-Jackson, Proc. Roy. Soc. (London) **B146**, 270 (1957); reproduced from originals kindly supplied by the author]. ×100 000.

The first experiments were carried out by exposing such reconstituted 640-A axial-repeat collagen fibrils to calcium-phosphate (Ca-P) solutions shown previously to be metastable with respect to the formation of apatite crystals. Because of the technical difficulties involved, and because the thermodynamic properties of calcium-phosphate solutions and their relation to the formation of solid phases are not well-enough characterized, it has not yet been possible to procure quantitative data on nucleation rates, crystal growth rates, free energy of nucleation, etc. The more qualitative methods used were designed to detect the ability of the test materials to *induce* the formation of apatite crystals, and depended mainly on the identification of the formed crystals. They did not distinguish between nucleation rate, crystal growth, recrystallization, etc.

The results of these experiments demonstrated that native-type, 640-A axial-repeat, reconstituted collagen fibrils, prepared from normally uncalcified tissues such as rat-tail tendon, calf skin, guinea-pig skin, fish swim bladder, etc., were able to nucleate apatite crystals from metastable calcium-phosphate solutions. Figure 15 is an x-ray diffraction pattern of such a calcified collagen preparation showing the typical diffraction rings of apatite.

Ironically, we were not able to prepare reconstituted collagens from bone, presumably because of its highly crosslinked nature. However, bone could be decalcified under a variety of conditions, which maintained the 640-A structure of many of the collagen fibrils. This

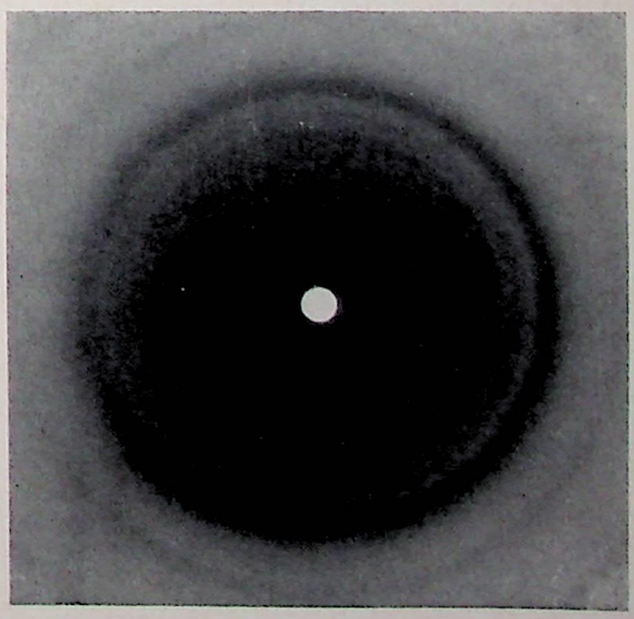


FIG. 15. X-ray diffraction pattern of *in vitro* calcified collagen. Note the lack of crystal orientation (evidenced by complete rings) and the broadness of the diffraction lines attributable primarily to small crystal size.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

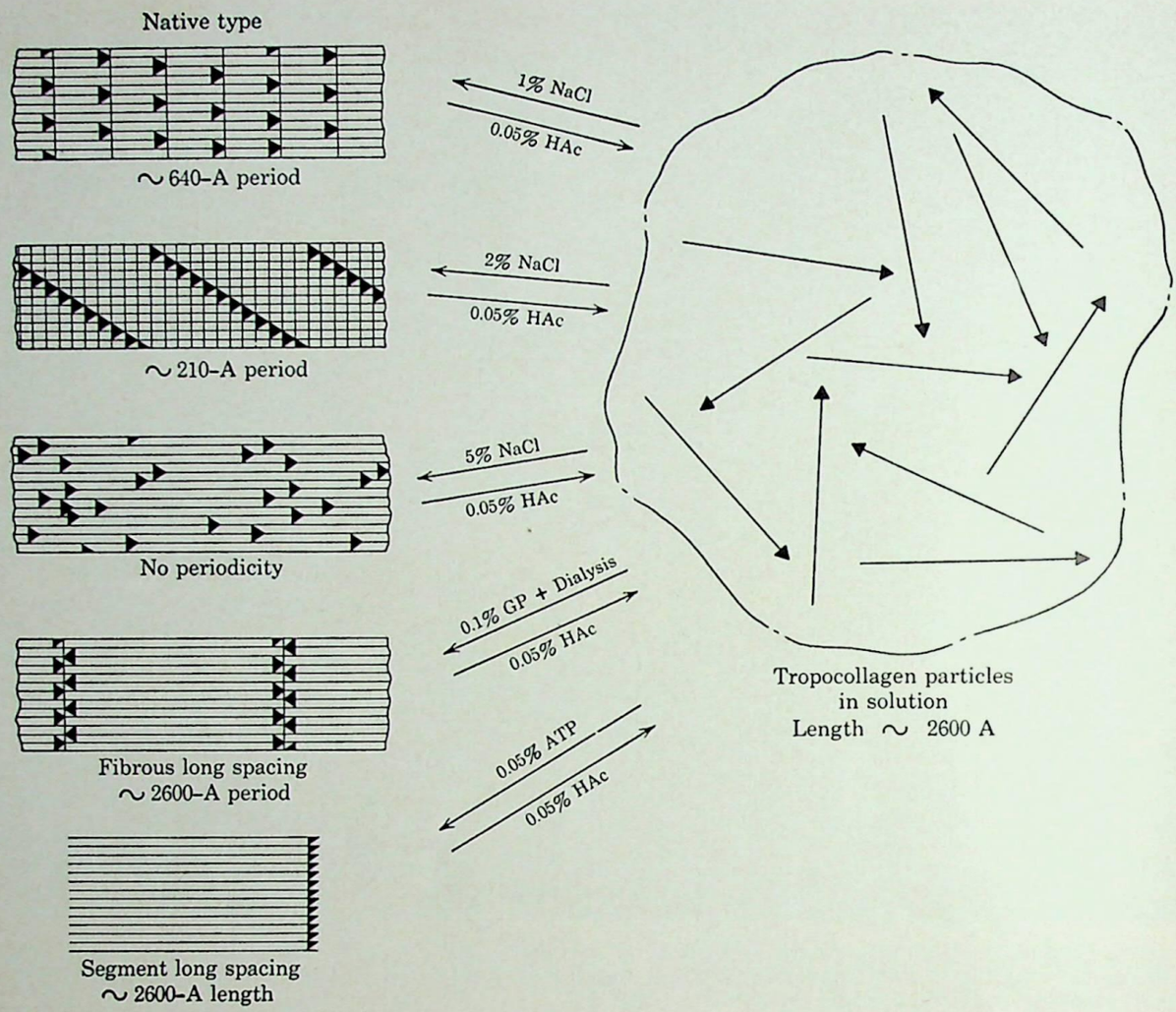


FIG. 16. Modes of aggregation of collagen macromolecules *in vitro* [from F. O. Schmitt, Proc. Am. Phil. Soc. **100**, 476 (1956)].

preparation was also able to induce crystallization of apatite. Although it is very difficult to make quantitative comparisons, it appeared that many of the decalcified bones were not grossly as potent in this respect as the reconstituted collagens, presumably because of changes in the structure of many fibrils and to alteration of certain important functional groups during decalcification.[49–51]

Since, in nucleation phenomena, a number of discontinuities are able to induce phase changes, the next problem was to determine whether or not this property was specific for collagen and, if so, where the specificity resided. Paramyosin fibrils, another well-organized and ordered fibrous protein structure, showing a regular periodic pattern by electron microscopy and low-angle x-ray diffraction, were obtained from the adductor muscles of clams, and used in similar experiments under identical physiochemical conditions. They were not able to induce this phase change.

A number of other fibrillar forms can be reconstituted *in vitro* from the same solution of collagen macromolecules (see F. O. Schmitt, p. 349). These include fibrils with 220-A axial periods, structureless fibrils with no discernible band pattern, fibrous long-spacing (FLS), and segment long-spacing (SLS) (Fig. 16). These are *different* ordered-aggregation states of the *same* macromolecules and reflect differences in the packing and intermolecular geometry of the fibrils. These differences are, of course, also reflected in differences in the *kinds* of amino-acid side chains which interact, and in differences in their *stereochemical relations* to one another. In addition, the isolation of the collagen macromolecules and linear polymers of collagen macromolecules (protofibrils), both in solution and in the solid state, provided a unique system with which to determine the steric nature of the specificity.

Figure 17 is a schematic representation of the basis for these experiments. Using SLS as a typical example with which to compare the native-type fibril, it is apparent that within any region of the fibril, although the *same* macromolecules and, therefore, the *same* amino acids are present, the *steric* relations between amino-acid side-chain groups are different.

The macromolecules, protofibrils, and the various aggregation states were exposed to metastable Ca-P solutions. Neither the macromolecules, linear aggregation of macromolecules, nor any of the fibril forms, *other than those with the native-type 640-A axial repeat,*

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

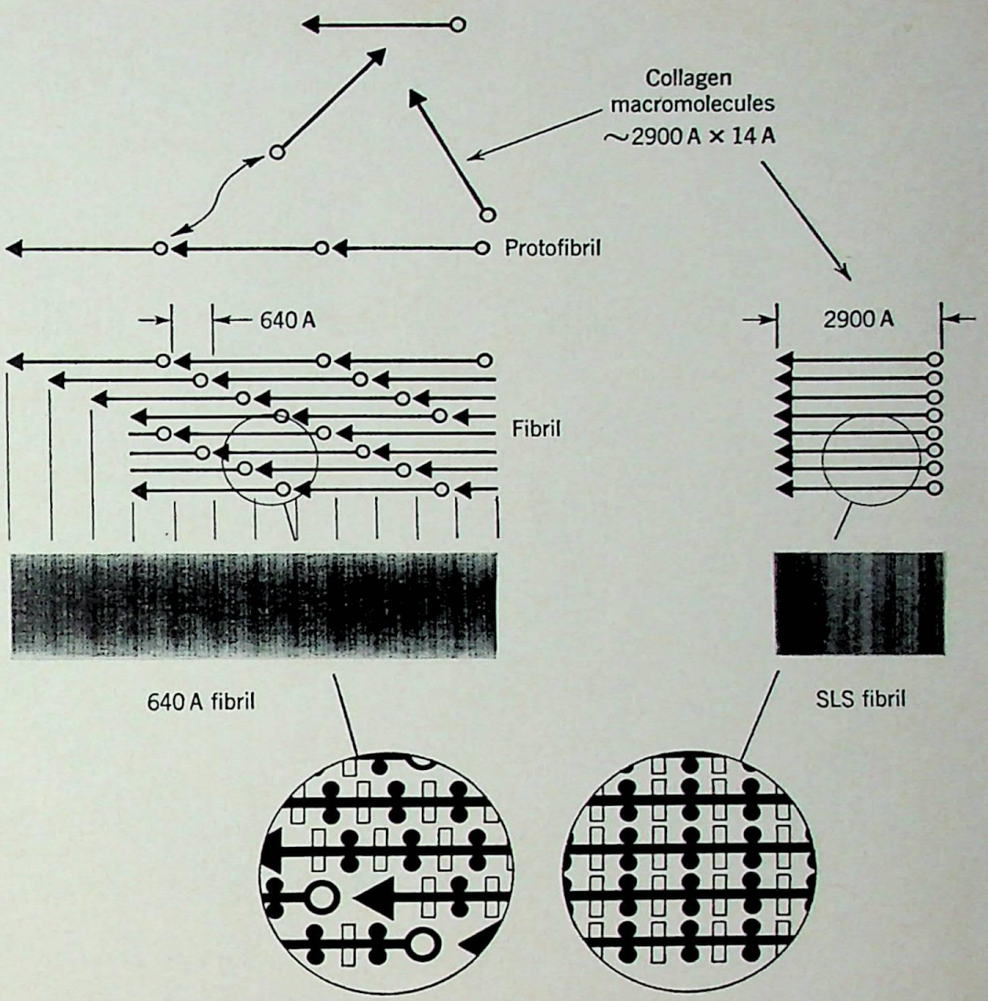


FIG. 17. The steric differences between the reactive side-chain groups of adjacent macromolecules in the native and SLS forms of collagen diagrammatically portrayed.

were able to initiate this phase change under identical physicochemical conditions. Since it was possible to pass reversibly from one fibril form to another and to demonstrate nucleating ability *only* while in the native 640-A form, it was evident that failure to initiate crystal formation was the result of the unfavorable configurations of the other types of reconstituted collagen fibrils and not the result of denaturation or other changes in the macromolecules produced in aggregating them from solution, or of the nature of the metastable solutions.

These experimental findings indicated that the process of *induction of crystallization* was a heterogeneous nucleation of apatite crystals from metastable Ca-P solutions by the collagen fibrils, and was *not* dependent on the macromolecules *per se* or on single linear adlineations of macromolecules, *but on groups of macromolecules and protofibrils polymerized laterally and longitudinally in a highly specific fashion characteristic of native collagen.* Thus, it would appear that the particular type of aggregation and packing of macromolecules characteristic of native-type fibrils creates highly specific steric relationships between reactive amino-acid side-chain groups from adjacent macromolecules within the fibril which serve as centers for nucleation.

Additional confirmation of the necessity of a *specific juxtaposition* of *certain* reactive *groups* was obtained by subjecting aliquots of reconstituted native-type collagen fibrils and demineralized bone to a number of physical and chemical agents, without altering the reactivity of the amino-acid side chains, and demonstrating that the ability to induce crystallization had been lost.

These included the effect of heat, acid, and alkali. In the case of heat, no change was noted *until* the thermal-shrinkage temperature was reached. Thermal shrinkage of collagen results in a phase change as regards its state of aggregation similar to gelatinization.[52,53] This disrupts both the molecular structure and the macromolecular-aggregation state of collagen.

In the case of acid-treated demineralized bone, or alkaline-treated reconstituted fibrils, although the collagen fibrils undergo a good deal of swelling and distortion, leading to loss of the typical 640-A repeat by low-angle diffraction and electron microscopy, the *macromolecular structure* remains intact. In both cases, the amino-acid side chains are not *chemically* altered, but only their *steric* interrelations are changed.

With reference to the phenomenon of nucleation, three distinctly different aspects of the relation between collagen and apatite must be considered. The first is the stereochemical relationship between the reactive side-chain groups which constitute a nucleation center. The second is the demonstration that there are *preferred*

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

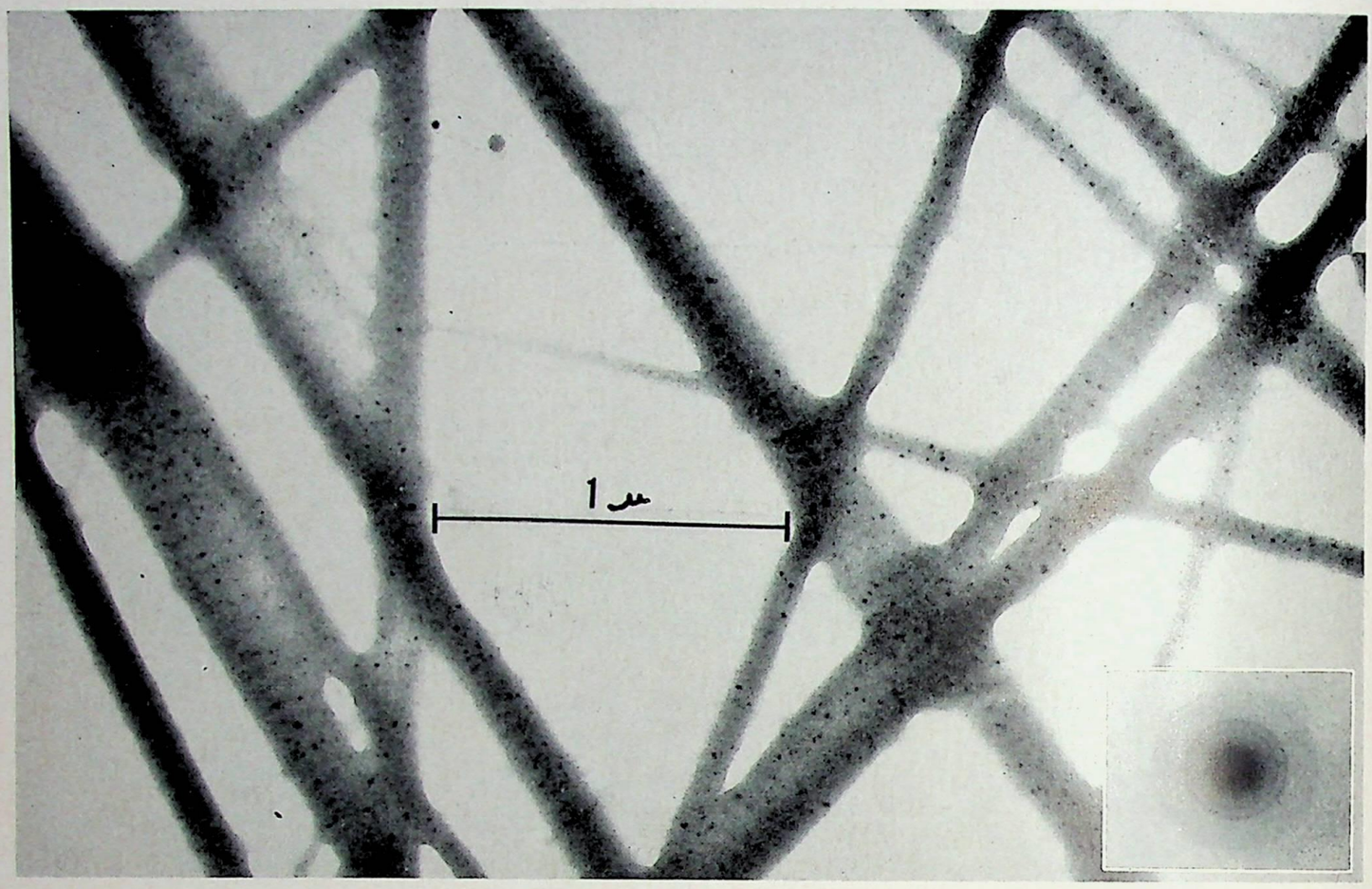


FIG. 18. Unstained, unshadowed preparation of an early stage of *in vitro* calcification of collagen fibrils. Note the regular and periodic distribution of the crystals along the collagen-fibril axis corresponding to the intraperiod fine structure of the fibrils and occurring primarily once per axial period.

FIG. 19 (lower right-hand corner). Selected-area electron-diffraction pattern of preparation similar to Fig. 18, showing the characteristic apatite reflections. There is no evidence of preferred crystal orientation, even though the diffraction pattern was obtained from an area where the fibrils were well oriented. ×42 000.

nucleation centers and the determination of their location with respect to the known electron-microscopic intraperiod fine structure of the collagen fibrils. The third is the nature of the chemical groups in the nucleation centers, and the nature of the intermolecular forces between these groups and the mineral ions.

An undue emphasis seems to have been placed on the *absolute* value of *640 A*, characteristic of native-type fibrils, and its relation to nucleation and other calcification phenomena. It should be apparent, however, that there is nothing specific about this *absolute* value, since it is merely a visible manifestation of the particular aggregation state of the collagen macromolecules and acts as a "fingerprint" for the recognition of the native-type fibrils. Its importance lies in the fact that when the macromolecules are arranged in a specific three-dimensional array so that this characteristic electron-density distribution occurs along the fibril, *certain regions* are created within the fibril whose reactivity and steric relations permit them to act as sites of heterogeneous nucleation. It is this stereochemistry *between reactive side-chain groups within the nucleating regions* which is directly related to the mechanism of nucleation, and this has no direct relation to the *absolute* value of 640 A.

In attempting to correlate the molecular structure of collagen and the property of nucleation, therefore, one must distinguish clearly between the structural characteristics of the *macromolecules* and the higher-ordered configurations which are the result of the specific *arrangement* of the *macromolecules* in the *fibril*.

Unfortunately, unlike the case of nucleation of inorganic crystals by other inorganic crystals, a direct correlation cannot be made between the wide-angle diffraction spacings of collagen and the known lattice spacings of apatite, as has been attempted by one group of workers.[16] The wide-angle diffraction pattern of collagen cannot be treated as arising from simple Bragg planes as in inorganic crystals, but must be interpreted on the basis of the theory of helical diffractors.[54,55] Except for the 10- to 12-A equatorial spacings, indicating the distance between the centers of adjacent macromolecules, the wide-angle reflections are characteristic of the triple-chain helical-coiled structure of the *macromolecule* and give *no* information as to the interatomic distances and configurations of the side-chain groups in the *fibril, which* information *is* directly related to the nucleation phenomena. It should be clearly understood that not only do the 640-A-type

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

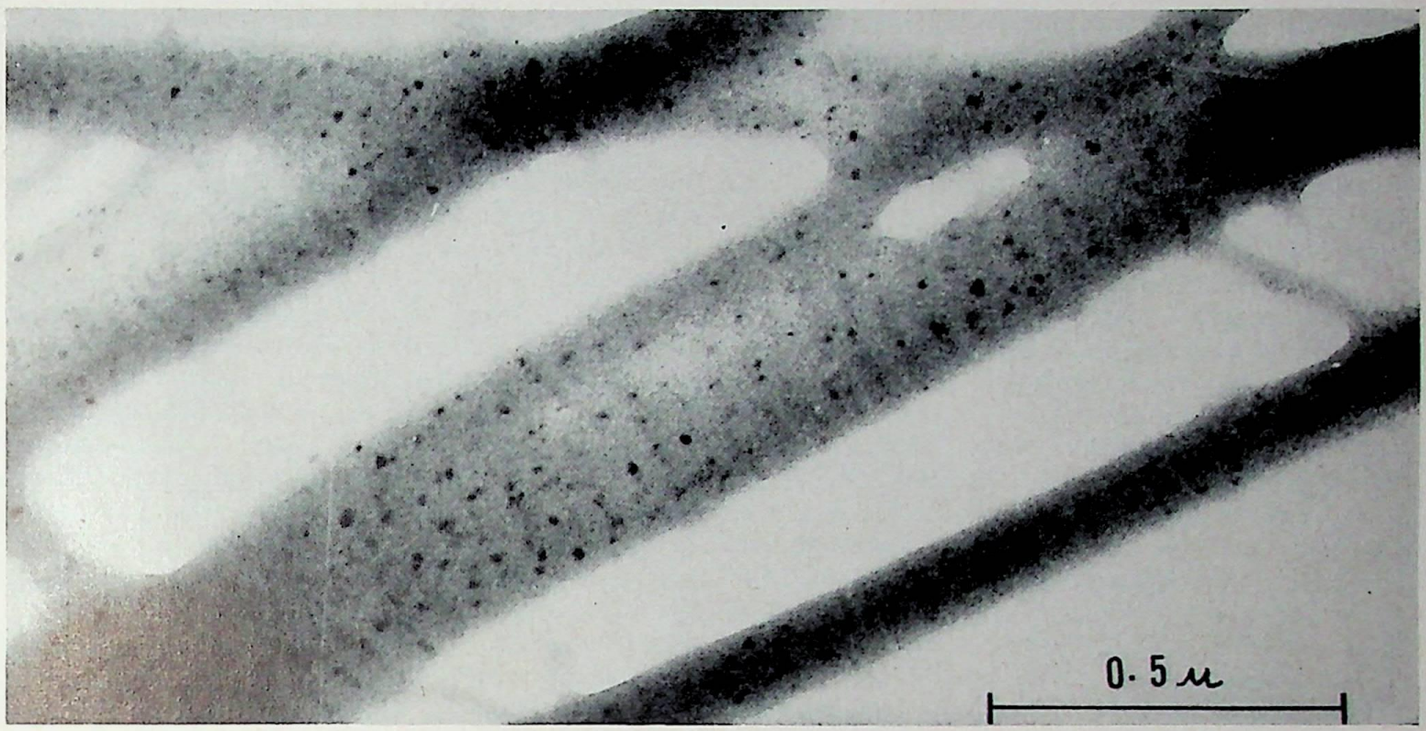


FIG. 20. Higher magnification of an area in Fig. 18. ×80 000.

fibrils give the characteristic wide-angle x-ray diffraction pattern, but also the macromolecules themselves and all other fibrillar forms, as well as cold gelatin films.

To date, neither the exact sequence nor stereochemistry of the amino acids in the collagen macromolecule are well enough known to permit determination of this specific configuration, although some general statements may be surmised from other data presented below.

As regards the second point—the demonstration of specific nucleation centers and their location—we have attempted to elucidate this aspect by several different methods. Direct visualization by electron microscopy and correlation of the changes in electron density, either in electron micrographs or from low-angle x-ray diffraction patterns, have been tried.[56] Interpretations based on electron-micrograph density changes and low-angle x-ray diffraction density changes are fraught with technical and theoretical difficulties, and no definite statement is as yet possible from these data.

Electron micrographs taken during the course of time-sequence studies of *in vitro* calcification (Figs. 18–20) and during the earliest stages of *in vivo* calcification of embryonic bone (Figs. 13, 14) provide direct evidence that there are specific regions in the fibrils which act as nucleation centers. These show that, during the earliest stages of calcification *in vitro* and *in vivo*, small dense particles, varying in size from 20 to 150 A, are deposited *regularly spaced* along the fibrils. Selected-area electron-diffraction of these particles in both cases revealed that they are crystals of apatite (Fig. 20).

Since the smallest crystals visualized in electron micrographs represent the *summation* of nucleation and crystal growth, interpretations based on such visual observations must be viewed with some caution. It is entirely possible that regions most favorable for crystal growth may be quite different from those where nucleation occurs. Also, there may be different areas capable of nucleation but differing in their degree of catalytic potency.

In addition, it should be pointed out that the nucleating center is not represented by an entire "band" or "interband" in the collagen fibrils, but only by *regions* within such transverse sections of the fibril. This can be deduced easily if one considers that it takes a certain number of repeating groups from adjacent macromolecules in just the right configuration to make a nucleation center, and that statistically there will only be a finite number of such groups with the proper steric configurations within any one transverse section.

Although it has not been possible to identify the exact location of the nucleation centers with respect to the intraperiod fine structure of the collagen fibrils, observations thus far indicate that these centers do correspond to certain of the so-called band (electron dense) regions. This would correlate well with related chemical and low-angle diffraction studies[57,58] which have shown that the band regions are the most reactive sites for many electron stains, tanning agents, etc., and that these areas probably contain many of the amino acids with long, reactive, polar side-chain groups.

In this respect, it is interesting to note the recent demonstration of the direct correlation between the degree of mineralization and the reactivity of the ε-amino groups of lysine in bone and tooth during demineralization,[50] since lysine is considered to be a major constituent of the band regions.[57–60]

Whereas these experiments have indicated the *steric*

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

specificity in the heterogeneous nucleation of apatite crystals by native-type collagen fibrils, they do not provide information concerning the manner in which these steric factors operate or concerning the nature and importance of the chemical interaction between collagen and the mineral ions, and the specific amino acids and mineral ions involved.

In this regard, certain general principles may be stated. Thus, while it is obvious that some sort of chemical interaction must occur between certain reactive groups in the collagen fibrils and the appropriate ions for nucleation to occur, it is imperative that the nature of the forces involved be such that the ions are still capable of interacting with the other constituent ions of the crystal lattice. For, if either calcium or phosphate ions were very *strongly bound* by the collagen fibrils, collagen would act as a *demineralizer* similar to chelating agents, unless a second mechanism were invoked which *released* the ions after they were bound.

There are several distinct ways in which the steric factors could operate to induce crystallization. The simplest case would be by facilitating the local concentrations of calcium and/or phosphate ions without requiring that the ions themselves be arranged in any particular steric fashion. This local increase in ion concentration would lead to crystallization by exceeding the metastable limit.

Another possibility is that the precise array of groups necessary for nucleation imparts a specificity for the *selective adsorption* or binding of the calcium or phosphate ions, either randomly or in specific steriochemical fashion, and that they then interact with the other constituent ions of the lattice to produce the first fragments of the new phase.

A third method would be the most specific of all, and would statistically require the fewest number of atoms. In this case, the precise array of reactive groups would sterically closely approximate certain low-index planes of apatite in a fashion similar to that proposed by Turnbull *et al.*,[61–63] and discussed in the Appendix. Thus, calcium, phosphate, and hydroxyl ions, either singly or in combination, depending upon which atomic plane of apatite is involved, would be "lightly" bound in such a *geometric configuration* that they would constitute a reactive nucleus.

Of course, one must consider the possibility that certain *clusters* of ions are bound by the reactive groups in the collagen structure. This is really a semantic issue, however, since the binding of such clusters, which presumably have the configuration of the bulk-solid phase, would require similar steric considerations.

Experiments in which nucleation failed to occur, after native-type reconstituted fibrils and demineralized bone were alternately exposed many times to solutions containing *either* calcium or phosphate ions in concentrations equal to or several times greater than those of metastable solutions from which nucleation readily occurred, confirmed the supposition that the intermolecular forces between collagen and the inorganic ions must be just strong enough to affect their interaction energies without *firmly* binding them to the collagen structure.

As to the specific organic groups and mineral ions involved in calcification, most investigators have felt that the initial step in calcification involves the presence in the organic matrix of an anionic group which combines with calcium ions. The anionic site has been variously assigned to chondroitin sulfate,[37] to chondroitin sulfate-collagen complexes,[35,36] or to a phosphorylated polysaccharide.[64] In addition to the fact that the experimental evidence indicates that the nucleating centers consist of reactive amino-acid side-chain groups in the collagen fibril, other considerations cast some doubt on the hypothesis of the *primary* role of calcium-binding in crystal induction.

In the first place, the apatites are *phosphate* salts, and the structural characteristics of the apatite lattice are primarily attributable to the phosphate groups, not to the calcium atoms which can be replaced by a number of other cations (Sr, Pb, etc.) without changing the major features of the crystal structure and symmetry. Since the phosphate groups are the "backbone" of the lattice, their role in the formation of the initial crystal structure would appear to be equally as important, if not more important, than that of the calcium ions.

The findings of Solomons *et al.*,[50] referred to earlier, which showed a direct correlation of the available ϵ-amino groups of lysine (and hydroxylysine) and the degree of mineralization in bone and tooth also is very suggestive, and may indicate that the primary collagen-mineral-ion interaction is between the ϵ-amino groups of lysine and the phosphate ions. Phosphorylation of NH_2 groups has been proposed also by Polonovski and Cartier[65] as the initial step in calcification. Experiments, now under way in this laboratory, in which specific amino-acid groups are being blocked singly and in combination, should provide the necessary data for the interpretation and elucidation of the actual molecular mechanism of the nucleation process.

LOCALIZATION, REGULATION, AND INHIBITION OF CRYSTALLIZATION

The phenomena of *localization*, *regulation*, and *inhibition* of the physicochemical mechanism which initiates crystallization both in normally mineralized and unmineralized tissues are different but closely related. On the basis of the hypotheses and experimental data presented, it would appear that all collagenous matrices are *inherently capable* of nucleating apatite crystals from metastable solutions. Since collagen is the major fibrous protein of all of the connective tissues (skin, tendon, ligaments, etc.), the questions arise: Normally, why do all of these tissues *not* calcify? Why *do* they calcify in certain pathological conditions? Also, even under normal circumstances, apatite crystals are deposited in tissues which do *not* contain collagen fibrils (enamel),

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

and in abnormal states in other noncollagenous tissues as well.

It is impossible to answer all these questions definitively at the present time, but the controlling factors and the circumstances under which they may be operative can be developed within the framework of the physicochemical concepts of solution metastability and heterogeneous nucleation sites.

Localization of the Crystals to Specialized Tissues

It is not suggested that collagen is the *only* organic compound capable of nucleating apatite crystals. This is quite unlikely, as, in all nucleation processes, many materials can act as nucleation catalysts with varying degrees of potency. In the case of tooth enamel, the apatite crystals are closely associated with another structural and "crystalline" protein, a eukeratin,[66–68] and it is likely that a similar mechanism for crystal induction exists here.

Regulation and Inhibition

Since the native collagen fibrils of most connective tissues do *not* calcify under normal conditions, one of several general situations or a combination of them must exist. Either the collagen in those tissues *not* normally mineralized is different from that of bone or calcified cartilage; or the degree of metastability of the extracellular fluids immediately surrounding and within the collagen fibrils in the various tissues is different; or other local phenomena increase the catalytic potency of the collagen fibrils in the normally mineralized tissues.

Since *reconstituted* native-type collagen fibrils from a wide variety of tissues normally *not* calcified *were* able to initiate crystal formation *in vitro* in our experiments, *failure* to mineralize, if attributed to the collagenous component, would involve subtle *structural* differences between native and reconstituted fibrils, as yet unresolved by the physical and chemical methods employed to date. However, it is possible that, during the extraction and reconstitution of the fibrils, parts are lost of the collagen macromolecules which normally inhibit calcification *in vivo*. There are no data available with which to evaluate the feasibility of this suggestion.

With respect to the second possibility, that differences in the degree of solution metastability account for the specific localization and regulation of calcification, two different points of view are possible. If one assumes that the degree of metastability of the unaltered interstitial fluid is sufficiently great so that collagen *can* induce perceptible rates of nucleation, some mechanism must be operative in the normally unmineralized tissues which *prevents* crystallization. If the degree of metastability of the unaltered interstitial fluids is *not* sufficiently high, either an *increase* in the degree of metastability or an increase in the catalytic potency of the collagen fibril is necessary for crystallization even in the normally mineralized areas, and a *minimal* protective mechanism is required in the normally uncalcified regions.

These variations in the degree of metastability could result either (a) from cellular-controlled compositional changes in the interstitial fluids, or (b) from the presence of other substances in the tissue which actively or passively controlled diffusion and specific ion transport and transfer, or competed with the mineral ions for active sites in the collagen fibril.

Cellular-controlled compositional changes of the interstitial fluids include variations in the calcium and/or phosphate concentrations, calcium to phosphate ratios, *p*H, ionic strength, ion complexes, etc. These could be mediated directly by the cells or by substances secreted by the cells, such as enzymes. For example, the demonstration that phosphorylative glycogenolysis can produce a local increase in phosphate concentration in epiphyseal cartilage[34] illustrates the role of such a device in the *regulation* of calcification in *normally mineralized tissues*, either by merely making more phosphate available, or by participating in another enzymatic cycle[69] which actively transfers phosphate ions to specific groups in the collagen fibril.

As to another method of *regulating* the degree of metastability (particularly in *normally unmineralized tissues*) by controlling diffusion, ion transport etc., the possible role of the ground substance and certain of its components have also been investigated.

Although much research has been done on the ground substance of connective tissues, its exact composition, state of aggregation, and anatomical distribution are not clear. Prominent among its components are the various acid mucopolysaccharides. These are thought to exist in the tissues as complexes with noncollagenous proteins.[70,71] In the past, most investigators have postulated that, of the ground-substance components, the chondroitin sulfates specifically, either alone or in combination with collagen, have played a role in the initiation of calcification.[35–37]

That it is not solely the *amount* of the acid mucopolysaccharides which determines whether calcification is initiated is obvious from the fact that hyaline cartilage —one of the richest sources of this material—is normally uncalcified, whereas bone, which does calcify, contains extremely small amounts of these substances.

Differences in the *kinds* of mucopolysaccharides present is not a plausible explanation, since the various acid mucopolysaccharides found in adult bone or in the growing ends of bone (bone, epiphyseal cartilage, etc.) are present in other tissues.[72]

It is instructive to review the findings as to the nature of the ground substance in cartilage, since a transition occurs from the normally uncalcified hyaline portion to the normally calcified epiphyseal region over a relatively short anatomical distance.

When tissue sections of hyaline cartilage, epiphyseal cartilage, and bone are stained with metachromatic dyes or by the Hotchkiss procedure, definite differences

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

are noted.[43] The staining characteristics of hyaline cartilage gradually change as the epiphyseal cartilage is approached. This change is also evident in bone which is being actively deposited or resorbed. Since metachromatic dyes react with negatively charged, high molecular-weight compounds, these staining characteristics have been linked primarily to the sulfated mucopolysaccharides in such connective tissues. The change in metachromasia and in the staining properties of these tissues by the Hotchkiss procedure has been interpreted as evidence for the depolymerization of the anionic mucopolysaccharides, as the zone of calcification is approached and reached.[41–43]

Although there is obviously a number of other possible reasons for this change in staining properties, the important point is that *there is some alteration*, either in the amount, state of aggregation, or reactivity of charged groups, etc., in the ground substance, and that this change accompanies calcification. Analytical data[73] which demonstrate a very marked loss of organic sulfate during cartilage calcification and during the formation of bone matrix indicate that it is the *depolymerization* and subsequent *removal* of these compounds which are related to calcification.

In attempting to assign a specific role to the ground substance in calcification, two properties stand out as important. One is that its state of aggregation is probably that of a gel in which the mucopolysaccharides exist complexed with a protein moiety as a mucoprotein.[70,71] Such a physical state of aggregation might well limit the diffusion of interstitial fluids and ions to the collagen fibrils.

The anionic groups of this mucoprotein are free and reactive,[74] which accounts for the second property of importance—the large cation-binding capacity of the ground substance. Under normal circumstances, therefore, the mucoproteins may help to inhibit calcification by limiting the available mineral ions, both by diffusion and by selective cation (calcium) binding. It is also possible, but less likely (since collagen and the chondroitin sulfates, for example,[75] do not react with collagen above *p*H 4.0), that the reactive groups of the mucopolysaccharide portion of the mucoprotein compete with the mineral ions for positions in the collagen fibrils, but it is possible that the noncollagenous protein portion of the mucoprotein may interfere with the process by just such a mechanism.

Depolymerization and removal of these compounds by eliminating the diffusion barrier, decreasing the cation-binding capacity of the ground substance, and possibly "freeing" some of the reactive groups of the collagen from their interaction with the protein moiety of the mucoprotein, would allow the mineral ions to react with the collagen. The cation-binding capacity of the remaining depolymerized mucoprotein components would also most likely be decreased, since it has been shown that metallic cations are more strongly bound to high molecular-weight acids and bases than to their monomeric compounds.

In vivo, the rapid depolymerization of the mucopolysaccharide-protein complexes and the subsequent decrease in their cation-binding properties might well lead to a local release of free cations including calcium, so that the resultant *increase* in calcium-ion concentration might then actually *aid* the initiation of crystallization by increasing the degree of metastability and making additional calcium ions available.

This hypothesis as to the role of the ground substance is supported by a series of *in vitro* experiments, conducted in our laboratory, in which native collagen-rich tissues, such as rat-tail tendon, calf skin, guinea-pig skin, etc., *failed* to mineralize under physicochemical conditions identical with those used in testing *reconstituted* native-type fibrils *from these same tissues*. In addition, when these tissues were treated so as to *extract* many of the components of the ground substance, including the chondroitin sulfates, either directly or by enzymatic depolymerization, these same native tissues readily mineralized.

Further support for this hypothesis comes from the experimental observation that hyaline cartilage, rich in the acid mucopolysaccharide-protein complexes *fails* to mineralize *in vitro* but selectively *removes* Ca^{++} from the calcium-phosphate mineralizing solutions. This property has been shown to be related to the sulfate groups of the chondroitin sulfates.[77]

This "shielding" of reactive sites in the collagen fibrils by ground substance components in intact tissues has also been noted in the tanning and dying of skins, a process which depends on the interaction of certain dyes, complex metallic ions, etc., with specific groups in the collagen. It was found that such interactions were markedly facilitated when the tissues were first tested by procedures designed to extract the ground substance components, and even more so by the prior depolymerization of the chondroitin sulfates by testicular hyaluronidase.[78] The latter method, in addition to being more specific, was also carried out under milder conditions so that the collagen fibrils were presumably less distorted and less swollen.

Figure 21 is a schematic representation of many of the experiments described. It illustrates both the specificity of the collagen macromolecular-aggregation state in *initiating* calcification, and the *possible* role of the ground substance as *a factor* in inhibiting it.

Since nucleation rate is so markedly dependent on the degree of metastability (p. 390), the organism has the dual problem of keeping the metastability of the extracellular fluids at a sufficiently high level for collagen to induce a perceptible rate of nucleation in normally mineralized tissues, and at the same time not so high that aberrant calcification cannot be safely controlled. It seems likely that a compromise is obtained by maintaining the degree of metastability in unaltered extracellular fluid *just* below the point where collagen is very effective. The rate of nucleation in bone and cartilage could then be increased and controlled by very small increases in the local concentration and/or active

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

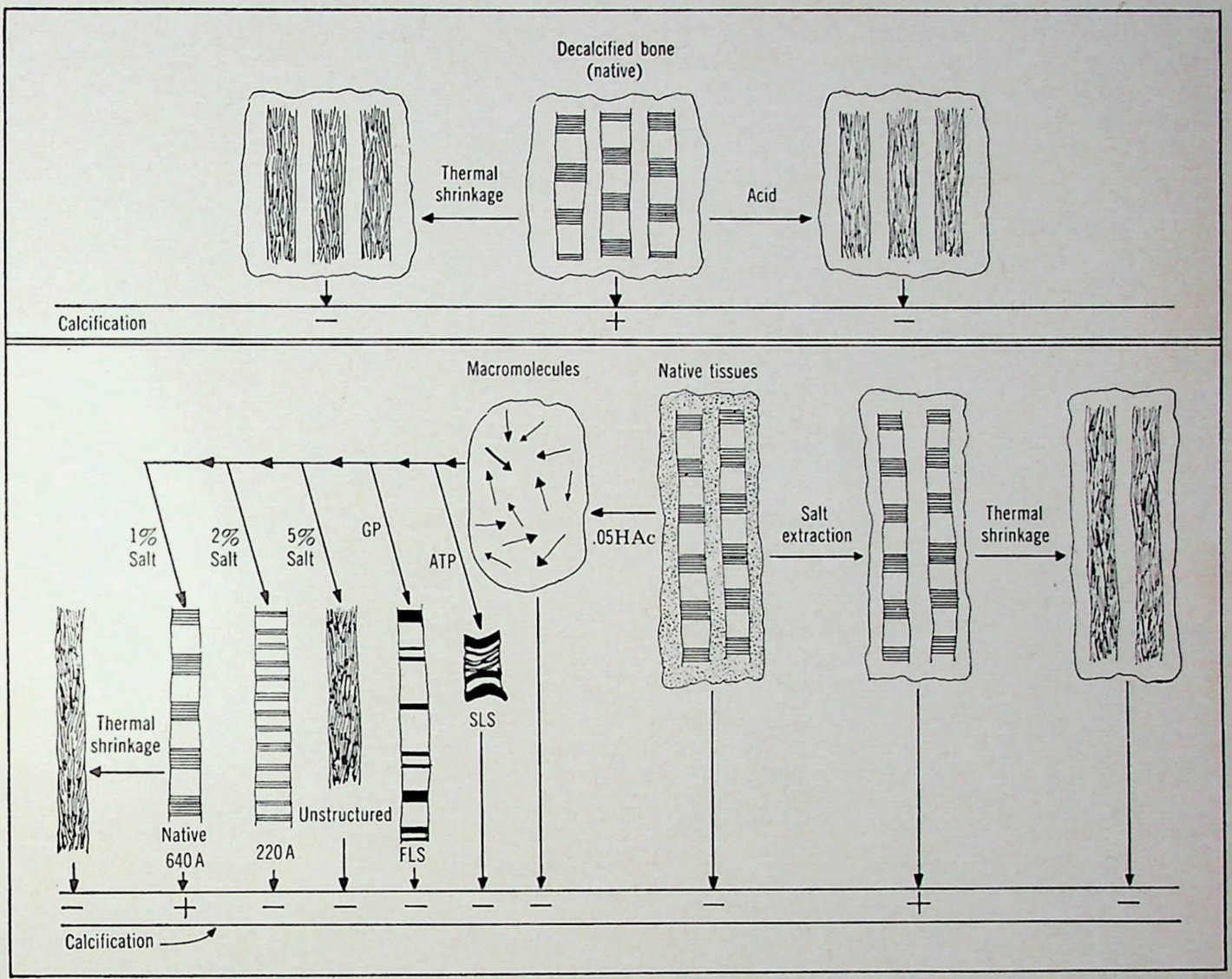


FIG. 21. Composite and diagrammatic illustration of the experiment demonstrating the specificity of the macromolecular aggregation state of native-type collagen fibrils in calcification, and the possible role of the ground substance (heavily stippled regions surrounding fibrils in native tissues) in inhibiting and controlling it. Enzymatic treatment of the native tissue is not shown, since it is not yet certain whether depolymerization is itself effective, or whether depolymerization and subsequent removal from the tissue is necessary.

transfer of mineral ions by enzyme mechanisms, such as the one mentioned earlier for phosphate ion, and the prevention of pathological calcification assured by decreasing the degree of metastability by utilizing the properties of the ground substance, as enumerated in the foregoing.

There is every reason to expect that, like other functions of such vital importance to the organism, mineralization is under the control of many factors, delicately balanced to provide *biological* and *cellular regulation* of the physiochemical *mechanism* which initiates crystallization.

COLLAGEN-INORGANIC CRYSTAL RELATIONSHIPS

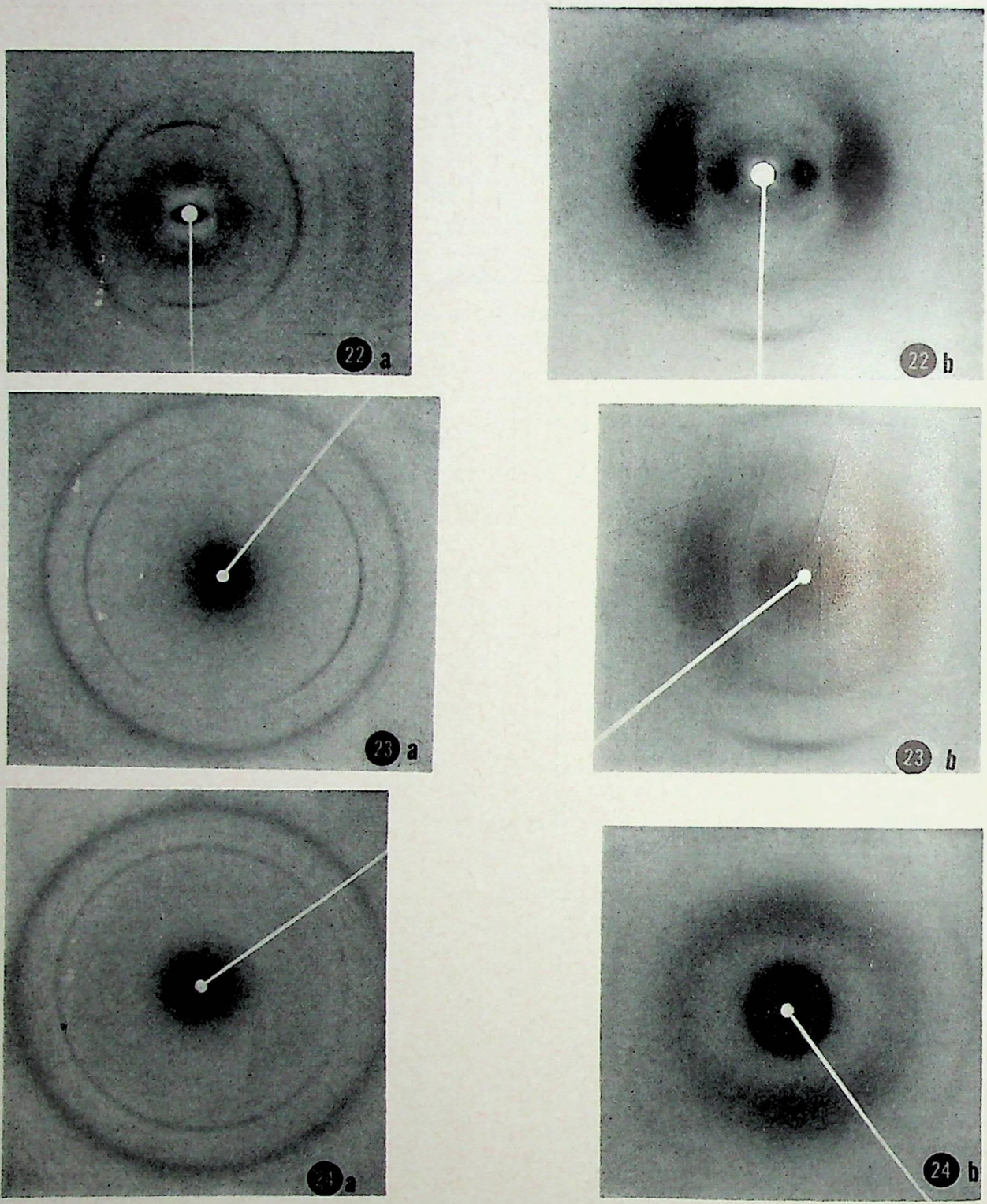
Location of the Inorganic Crystals with Respect to the Collagen Fibrils

The location of the crystals within the collagen fibrils supports the theory of heterogeneous nucleation as presented, since the probability of arranging a number of side-chain groups in the proper steric configuration is statistically higher *within* the fibrils, and where the density, intermolecular forces, and interaction energies are highest.

The longitudinal position of the crystals tends in many instances to be localized to certain regions accentuating the 640- to 700-A axial repeat of the collagen fibrils. It is difficult to say exactly where these regions are in relation to the intraperiod fine structure of the fibrils, although density considerations and shadowed preparations indicate that they are most likely in those regions where the density of the bands is greatest. This is especially true in the less heavily calcified fibrils. In many other fibrils, however, especially where calcification is quite high, no such localization is apparent. It would appear, therefore, that—although *initially* there are preferred regions in the fibril for nucleation and for crystal growth—as the fibril continues to calcify, crystals are deposited throughout the fibril.

X-ray diffraction evidence that the average size of the crystals is approximately 220 A led some workers to

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA



FIGS. 22, 23 and 24. X-ray diffraction patterns of longitudinally oriented native, calcified fish bone [Fig. 22(a)], rat femur [Fig. 23(a)], and embryonic metatarsal rudiment [Fig. 24(a)], showing apatite orientation; same specimens [Figs. 22(b)–24(b)] decalcified, showing collagen orientation.

speculate that three such crystals were aligned longitudinally per axial period of the collagen fibril.[79] Such an arrangement would result in the structure having an axial repeat of 220 A. This has not been borne out by direct visualization of the fibrils and the inorganic apatite crystals by electron microscopy.

Inorganic Crystal-Collagen Fibril Coorientation

Earlier workers have demonstrated by electron microscopy that the long dimension of the inorganic crystals, while closely parallel to the fibril axis in bone, is for the most part randomly oriented in calcifying cartilage.[80] Since the process of orientation was felt to be a true-oriented overgrowth directed by certain crystallographic planes in the collagen fibrils (epitaxy)[5,16] and directly related to the process of mineral phase induction, it was suggested that *possibly* different mechanisms were involved in the initiation of calcification in these two closely related tissues.[16]

This was based also on electron-microscopic evidence

CC-0. Panjab University Chandigarh Collection. Digitized by eGangotri Foundation USA

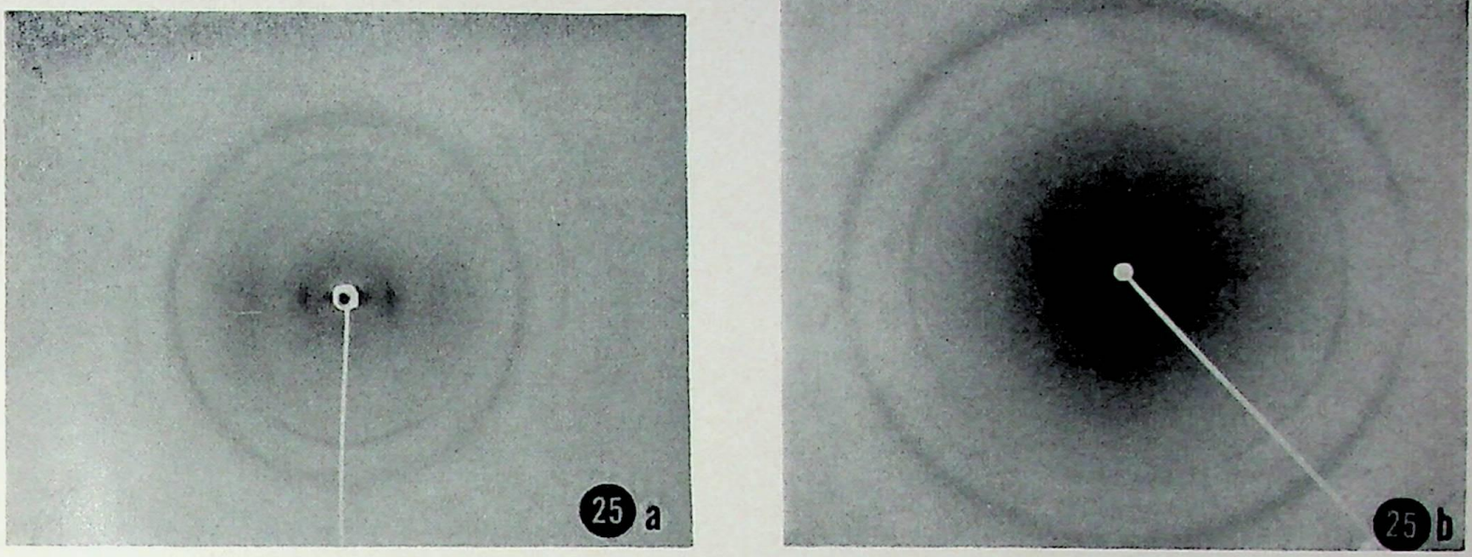


FIG. 25 X-ray diffraction patterns of *in vitro* (a) recalcified fish bone and (b) rat-femur bone. Note that, although there is *some* orientation of the apatite crystals as evidenced by the arcing of the 00*L* reflections, it is *not* as prominent as in the original native, calcified bone [compare Fig. 22(a) with Fig. 25(a), and Fig. 23(a) with Fig. 25(b)].

that, in the zone of provisional calcification in cartilage, the size, appearance, and organization of the collagen fibrils were markedly different from those in bone.[80] More specifically, the collagen fibrils in this calcifying zone of cartilage are quite thin (50 to 250 A) and do not show any visible interperiod fine structure. The fibrils are, in addition, widely separated and randomly oriented. Correlations of this type, however, depend both on an intimate knowledge of the *mechanism* of crystal orientation and its relation to the *induction* of crystallization, and on information concerning the macromolecular organization of the *apparently* "structureless" collagen fibrils. Furthermore, several recent studies also have demonstrated that the embryonic *bone* of some species (and even early postfetal bone in others) does not show any preferred orientation of the inorganic crystals[81–83] (as deduced by x-ray diffraction, based primarily on the arcing of the 00*L* reflections of apatite), and that crystal orientation varies not only from bone to bone, but also in different areas of the same bone.[82,83]

As in other problems in calcification, the term "orientation" must be defined carefully. Orientation of the inorganic crystals has been related to the long axis of the entire bone and to the various hierarchies of collagen structure. These definitions obviously are not identical, and the orientation of the crystals with respect to the *collagen fibrils with which they are associated* is the relationship of interest.

The evaluation and interpretation of such co-orientation by x-ray diffraction evidence alone is limited and may be quite misleading. The difficulties stem both from the basic nature of the method and from the failure to interpret the results in the light of the gross and microscopic arrangements of the collagen fibers in the tissue.

An x-ray beam, regardless of how small a collimator is used, integrates the orientation of the collagen and crystals over an enormous volume as compared with the order of magnitude discussed here. Since the x-ray diffraction data are summations from many crystals and fibers, it is not possible to discern *individual inorganic crystal-collagen fibril* relationships, particularly if the collagen fibrils are themselves randomly dispersed or if account is not taken of the higher-ordered organization of the structures in the tissue.

Electron microscopy, on the other hand, gives a direct result if the crystals can be visualized with respect to individual fibrils. In cases where small local areas are in question, selected-area electron diffraction—since it integrates over a much smaller volume than x-ray diffraction—also can be of great help.

In order to evaluate whether the marked variation in crystal orientation as deduced by x-ray diffraction is indicative of varying degrees of collagen fibril-inorganic crystal orientation, or whether it is the result of the geometric factors in tissue organization and of x-ray technique, a number of specimens showing the complete spectra of orientation was examined by electron microscopy, electron diffraction, and x-ray diffraction. Figures 22–24 are x-ray diffraction patterns of fish bone, rat bone, and 16-day-old embryonic chick bone, calcified and decalcified. In the case of fish bone, there is marked orientation of both the apatite crystals and the collagen fibrils. The rat bone shows less orientation of both the apatite crystals and the collagen fibrils, and in the embryonic bone, no orientation of the apatite is present and there is very little evidence of collagen orientation.

Electron micrographs and selected-area electron diffraction of the two extreme cases (fish bone and embryonic chick bone) clarified the issue. Figure 6 demonstrates that the fish bone consists of extremely well-oriented collagen fibrils with the inorganic crystals lying with their long axes (crystallographic *c*-axis) parallel to the collagen-fibril axis. Figure 26 is an electron micrograph of embryonic bone which reveals the general randomness of the collagen fibrils. Within very

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

FIG. 26. Electron micrograph of a section of the metatarsal rudiment of embryonic chick bone (16 days). Despite the general randomness of the collagen fibrils, the crystals are oriented in local regions corresponding to the individual collagen-fibril directions. The size and shape of the crystals are similar to those in adult bone. ×40 000.

FIG. 27. Selected-area electron diffraction of specimen shown in Fig. 26. The preferred orientation of the crystals is evident by the arcing of the 002 and 004 reflections.

local areas, however, *the same general parallel alignment of crystals and the collagen fibrils is present*, confirmed by selected-area electron diffraction (Fig. 27). The orientation of individual crystals with *the fibrils with which they are associated* is also apparent in certain adult bones where the collagen fibrils are clearly sepa-

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

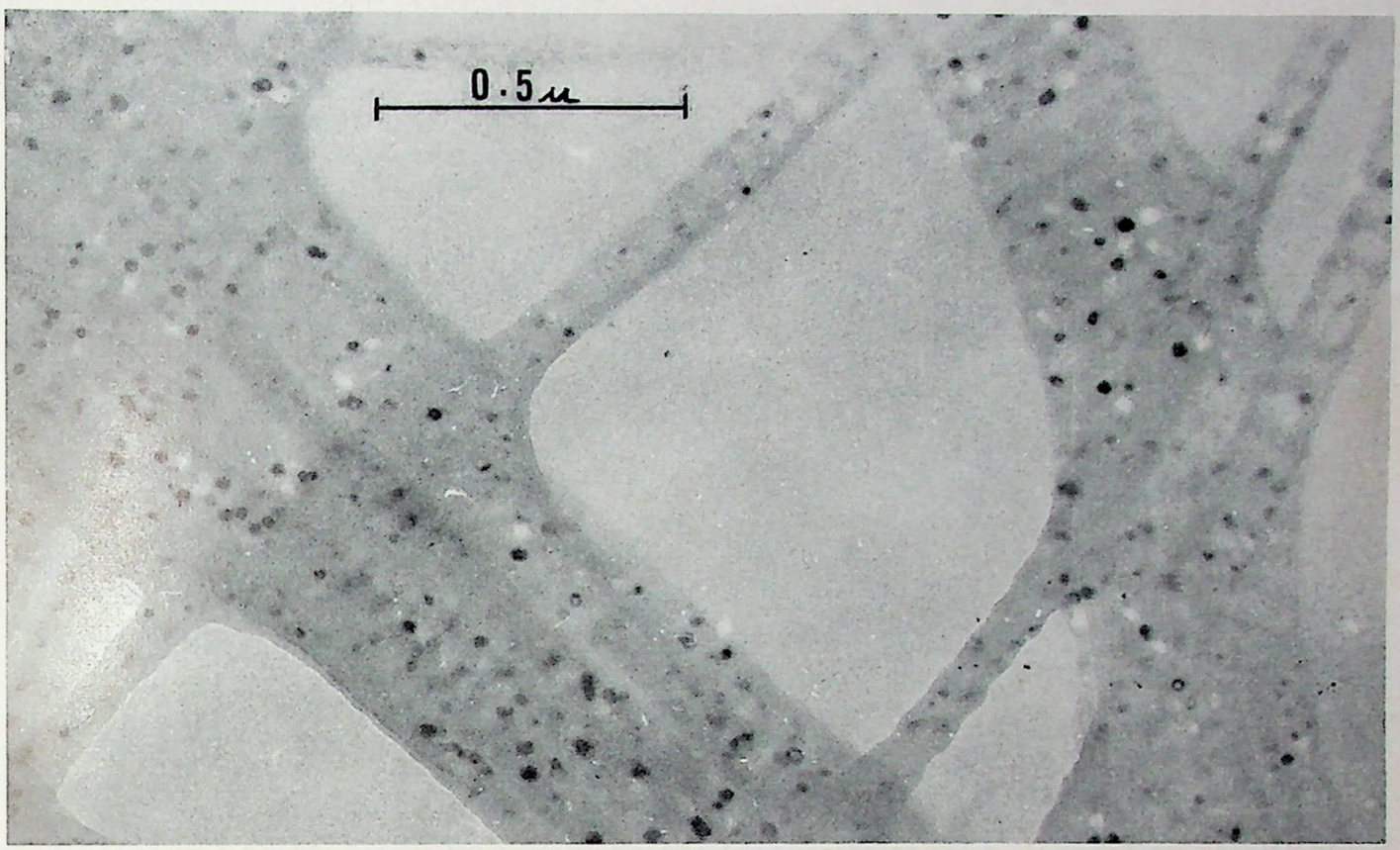


FIG. 28. Unstained, unshadowed preparations of *in vitro* calcified collagen at an early stage. Although the crystals are situated regularly spaced in certain regions of the fibrils, the haphazard arrangement of the crystal reflections (white spots) indicates lack of preferred *orientation* of individual crystals with the fibrils with which they are associated. ×70 000.

rated, as shown in Figs. 8 and 9. It is clearly evident, therefore, that x-ray diffraction evidence *alone* cannot be used as a criterion in assessing or interpreting crystal orientation, it may, as in this case, be quite misleading. Thus, the complete lack of preferred crystal orientation as evidenced by x-ray diffraction data in the embryonic bone is simply the result of the randomness of the collagen fibrils themselves, and, as far as the crystals within any one fibril with which they are associated are concerned, they are well oriented similar to adult bone.

Mechanism of Crystal Orientation

As further experimental results unfolded, it became apparent that the reasons for crystal orientation were twofold: *the size and habit of the inorganic crystals, and the location of the crystals within the fibrils.*

Electron micrographs taken during time-sequence studies of *in vitro* calcified reconstituted collagen have shown a striking similarity in the appearance of the inorganic crystals to those of embryonic bone during the initial stages of calcification (Figs. 13, 14, 18, and 20). In both cases, the crystals appeared dot-like with none of the axes elongated in any direction, and in both cases selected-area electron diffraction (Fig. 19) showed *no* preferred crystal orientation from areas where the collagen fibrils were relatively well oriented. On the other hand, in later stages of embryonic bone, once crystal growth (and/or recrystallization) has occurred, leading to *small* asymmetric crystals *within* the fibrils, the crystals do become oriented with their long axes approximately parallel to the collagen fibrils *with which they are associated* (Figs. 26 and 27).

Although it was impossible to obtain single-crystal patterns from the *in vitro* calcified specimens, individual diffraction spots, rather than complete rings, were obtained in many instances from small local areas with relatively few crystals and in regions where there were several well-oriented fibrils.

In order to obtain more information about the relationship of individual crystals to the collagen fibrils, recourse was had to another experimental device. The objective aperture of the electron microscope was removed and the specimen photographed slightly out of focus. The resultant crystal reflections, seen as white dots in Fig. 28, are from a definite set of crystal planes, confirmed by measuring the distance from the crystals to the reflections. One easily can see the haphazard patterns of these reflections, which clearly indicate that the individual crystals *are not oriented with their crystal axes parallel to the fibril axes, but are randomly oriented.* These data indicate that the *initial* crystals *in vivo* and *in vitro* have not grown in an epitatic fashion and that their *subsequent* orientation must be attributable to other factors.

A detailed electron-microscopic examination of calcification in cartilage has been made by Robinson *et al.*[80]

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

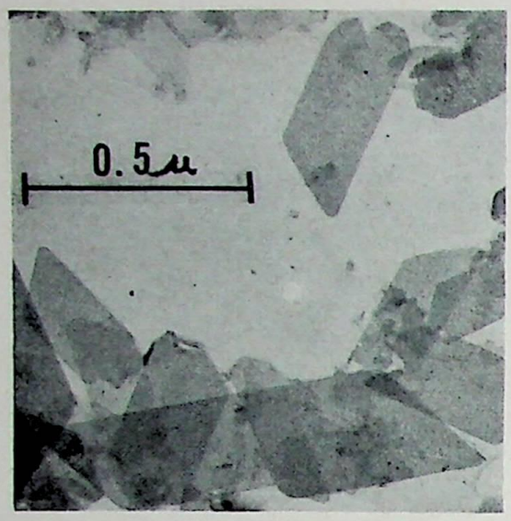


FIG. 29. Electron micrograph of *in vitro* apatite crystals precipitated under conditions similar to those of the *in vitro* collagen-nucleation studies. Note large hexagonal-type crystals. ×50 000.

They observed that, in the initial stages of cartilage calcification, the crystals *are* directly related to the collagen fibrils and are propagated in the intervening interfibrillar areas. The crystals are not preferentially oriented, except for a few which lie directly on, or possibly in, the fibrils.

The other point concerning calcifying cartilage—that is, the lack of the usual interperiod fine structure of the collagen fibrils—appears to be attributable to the very small size of the fibrils and possibly to the large amount of ground substance surrounding them. The question of collagen-fibril size is also important from the standpoint of the number of collagen macromolecules that must be polymerized laterally in order to constitute a nucleus.

With these factors in mind, reconstituted collagen fibrils were prepared, approximately 50 to 100 A in diameter and not showing cross-striations when stained with phosphotungstic acid (PTA). Well-oriented fibers of this material, however, demonstrated the characteristic 640-A low-angle x-ray diffraction pattern, indicating the *macromolecular organization* to be similar to the larger fibrils. Preparations of these fibrils were able to initiate crystallization of apatite *in vitro* from metastable solutions in experiments similar to those described.

The data may be summarized as follows. In calcifying cartilage, embryonic bone, and *in vitro* calcified collagens, crystallization is initiated in *direct relation to the collagen fibrils*. During the *first stages* of mineralization in embryonic bone, reconstituted collagens, and calcifying cartilage, the inorganic crystals are *not* oriented with respect to the collagen-*fibril axis*. In both embryonic bone and reconstituted collagens, the crystals are strikingly similar: dot-like with *none of the crystallographic axes elongated.*

Further crystal growth results in assymetric crystals with one of their axes elongated. In bone, this occurs primarily *within the fibrils* and is *associated with orientation of the crystals*. That the orientation of the crystals is not a function of the over-all compactness or organization of the collagen fibrils in the tissue, but is dependent upon the *location* of the crystals within the fibrils, is shown by the observations on embryonic bone and on certain adult bone. In the former (Fig. 26), despite the general randomness of the fibrils, and in the latter (Figs. 8 and 9), although the collagen fibrils are clearly separated and somewhat randomly directed, the inorganic crystals are well oriented with their long axes parallel *to the collagen fibril axis with which they are associated.* In calcifying cartilage, however, because of the very small diameter of the fibrils, crystal growth is propagated primarily *between* the widely spaced fibrils, and these crystals, therefore, *incapable* of being oriented by the fibrils, are randomly oriented.

The conclusions are reached that: (1) the basis for the orientations of the inorganic crystals is the asymmetric growth of small crystals *within the collagen fibrils, between tightly packed, longitudinally oriented chains of macromolecules* (*protofibrils*), which necessarily results in the crystals being similarly oriented; (2) the process of crystal orientation, therefore, is unrelated to the *mechanism of induction* of crystallization; (3) since crystal orientation is a function of the *size* and *shape* of the *crystals* and their position *within the fibrils* and not part of the crystal induction mechanism, the *lack* of preferred orientation in cartilage in no way suggests that the *mechanism* of crystal formation (heterogeneous nucleation by the collagen fibrils) is different in cartilage from that in bone.

In fact, the theory of heterogeneous nucleation of apatite crystals by collagen fibrils may be more convincing in cartilage than in bone, since the collagen fibrils are so closely packed (particularly in compact bone) that there is very little space for crystal formation *except* within the fibrils, whereas in cartilage, the fibrils are relatively far apart and randomly oriented with large amounts of intervening interfibrillar ground substance available for crystallization. Despite this, calcification does *not* begin randomly in the tissue, but is initiated in *direct relation* to the very thin fibrils and is then propagated throughout the intervening ground substance.

CRYSTAL HABIT AND SIZE

A good deal of disagreement exists as to the exact habit and size of the apatite crystals in bone. Electron micrographs obtained in our laboratory, in agreement with those of Speckman and Norris,[84] and Fernández-Morán and Engström,[85] show what *appear* to be rod-shaped particles, the majority varying in thickness from 15 to 30 A and in length from 200 to 400 A. It is possible, however, that these represent extremely thin, lathe-like crystals, several of which are stacked together. Perfect cross sections of well-oriented fibrils are obtained with great difficulty, but a number of fibrils in many fields appear to show a dot-like appearance in cross section consistent with a rod-like habit of the crystals (Fig. 12).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

Because the size of the crystals approaches the limit of resolution of the electron microscope, particularly in a tissue technically difficult for sectioning such as bone, and because of other factors such as overlap, etc., it is not possible to be absolutely certain of the exact size and shape of the inorganic crystals. In any event, the issue is somewhat semantic: when does a rod become a plate? There may be variations in the habit of the crystals from true rods to a *somewhat* lathe-like habit where one of the faces is *slightly* larger than the other. But from many cross sections, longitudinal sections and oblique sections, one can say with certainty that the crystals are *not* the large hexagonal plates (500×250×100 A) reported by some workers on the basis of the examination of blended or autoclaved bone.[27,28]

On the other hand, the apatite crystals formed by simple precipitation in the test tube,[86] or crystals nucleated and *grown in the presence of collagen*, are usually hexagonal plates quite unlike those occurring in bone.[87] Occasionally, lathe-like crystals can be formed in the test tube, but again these are several orders of magnitude larger than the bone crystals.[86] Thus, although the early crystals nucleated by collagen *in vitro* are similar to those nucleated *in vivo*, further crystal growth appears to be quite different.

This difference in crystal habit and size between bone crystals and precipitated calcium phosphate crystals might be explained on the basis of the mechanical factors resulting from crystal growth within the fibrils, or by an active process controlled by the structure of collagen such as has been suggested by others.[79] It is also possible that it is entirely unrelated to the structure of collagen.

The position of the bone crystals within the collagen fibrils suggests that the collagen fibrils play a primary role in determining crystal size and shape. The finding, however, that collagen fibrils were *not* able to alter the size or the habit of the crystals under the physicochemical conditions of the *in vitro* experiments casts some doubt on the validity of this hypothesis. This is also confirmed by an examination of electron micrographs of calcifying cartilage which reveals that the crystals lying in the interfibrillar space and *anatomically unrelated to the collagen fibrils* are the same size and shape as those within the collagen fibrils of bone.

Since it is a well-established fact that the size and shape of crystals can be altered by changes in the physicochemical environment—including the addition of a number of substances bound on certain crystal faces and limiting their further growth—studies were carried out on the precipitation and growth of apatite crystals from solution under a wide variety of experimental conditions.

Figure 29 is an electron micrograph of a calcium phosphate precipitate in which the physicochemical conditions were similar to the *in vitro* collagen-nucleation experiments. These large, hexagonal plates are somewhat larger than those usually seen under these circumstances, but serve as a control to illustrate the progressive change in size and habit that can be produced *in vitro* in the absence of collagen. X-ray diffraction of this preparation revealed a typical apatite pattern without any evidence of the presence of octocalcium phosphate or brushite which also crystallize in large plate-like form.

Figure 30 is another preparation in which an attempt was made to alter crystal size and shape. One can easily see the marked difference in the size of the crystals, although they are still in the form of hexagonal plates. Note that where the crystals overlap there is a suggestion of a dense "rod," particularly if viewed without reference to the neighboring crystals.

Figures 31–33 are electron micrographs of apatite preparations where the crystal size was further reduced, and it appears that the habit has also been altered so that a number of rod-like crystals have formed. It is impossible, however, to tell whether these are truly rods or represent extremely small, thin plates stacked together or supporting each other and viewed along their edges. The "rods" are now approximately the size of the bone crystals.

Figure 34 is an electron micrograph of another preparation in which the crystals are even thinner (many ~10 A or less) and very similar in size and appearance to the bone crystals, and in which there is no apparent evidence of hexagonal plates. If one examines the elec-

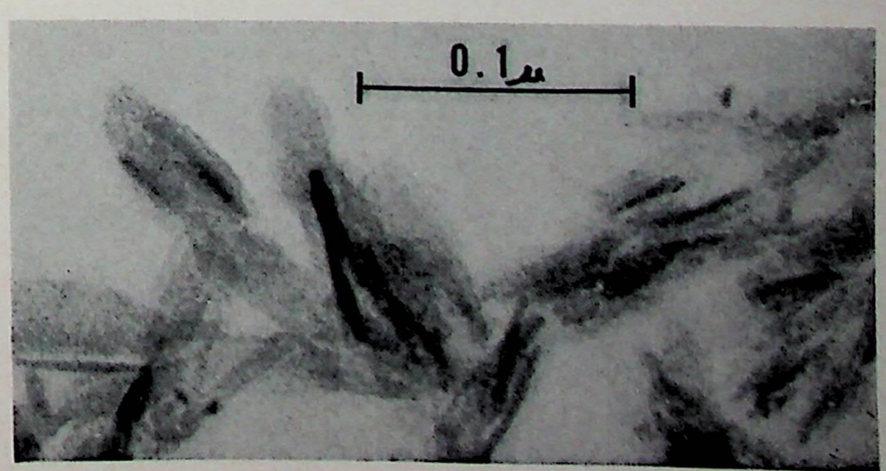


FIG. 30. Under altered physicochemical conditions, crystal size decreases markedly, but crystal habit remains the same. ×300 000.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

FIG. 31. Further alteration in the physicochemical environment produces not only a further decrease in crystal size but also, for some, a change in the crystal habit to a rod-like appearance. ×180 000.

tron micrographs carefully, however, it is obvious that there is a minimal but definite background electron density which may indicate that these crystals are extremely thin plates, some of which are stacked together or are upright and supporting each other, and, when viewed along their edges, give the appearance of rods. The fact that many of the rods appear to be bent in several directions may indicate that this is the correct interpretation. This same appearance has been seen in very lightly calcified collagen fibrils of bone.

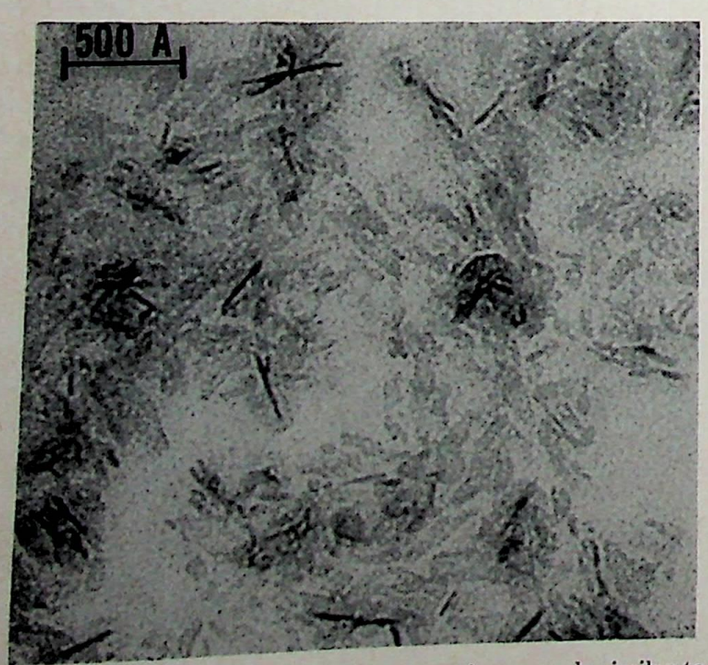


FIG. 32. Higher magnification of an area in a sample similar to Fig. 31, where the "plates" predominate. ×260 000.

Although it is still impossible to state dogmatically the exact size and shape not only of the bone crystals but also of many *in vitro* precipitated crystals, these experiments show that *it is possible to grow from solution under controlled physicochemical conditions apatite crystals whose size, shape, and appearance in electron micrographs and whose x-ray diffraction characteristics are similar to those of bone crystals.*

The facts (1) that *in vitro* collagen *per se* is not able to modify the crystal habit or size, (2) that *in vivo* (calcified cartilage) crystals similar in size and shape to the bone crystals lie in the interfibrillar spaces *between* the collagen fibrils and anatomically unrelated to them, and (3) that crystals *can* be grown *in vitro* similar to the size and habit of bone crystals in the absence of collagen, indicate that the *size* and *shape* of the *apatite crystals are primarily controlled by other physicochemical factors in vivo* and *not by the collagen fibrils.*

Obviously, the conditions under which crystal growth occurs within longitudinally oriented and closely packed fibrils may exert some secondary influence on the crystal size and habit, both mechanically and by selectively facilitating and inhibiting the diffusion of ions; but

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

when the other physicochemical conditions are not met, crystal growth is *not altered* by the collagen fibrils.

Under such conditions, the crystals ordinarily tend to grow to a relatively large size as compared with normal bone crystals, and since such crystal growth cannot take place within the fibrils, further seeding, recrystallization, etc., occurs *outside* the fibrils. The results of experiments in which crystals initially nucleated within the collagen fibrils *in vitro* proceeded to recrystallize and grow outside of the fibrils, substantiates this conclusion, and clarifies the observations related to crystal orientation in *in vitro* calcified collagens including bone.

Thus, *in vitro* calcified, reconstituted collagens showed *no* preferred inorganic crystal orientation by x-ray diffraction, even though the collagen was moderately oriented. On the other hand, specimens of bone recalcified *in vitro* did show some preferred inorganic crystal orientation, but not as much as similar samples of native bone, as shown in Figs. 22(a), 25(a) and 23(a),25(b). In the case of the recalcified bone, the degree of orientation of the crystals was related to the degree of orientation of the collagen matrix [Figs. 22(b) and 23(b)].

Electron micrographs of heavily calcified preparations of *in vitro* calcified collagens demonstrated that the majority of the crystals were in the interfibrillar space and *not* within the fibrils. In addition, the crystal habit and size were unlike that of native bone and resembled the large *in vitro* precipitated hexagonal plate-like crystals. Because of size and shape, crystal growth was more favorable in the interfibrillar regions rather than within

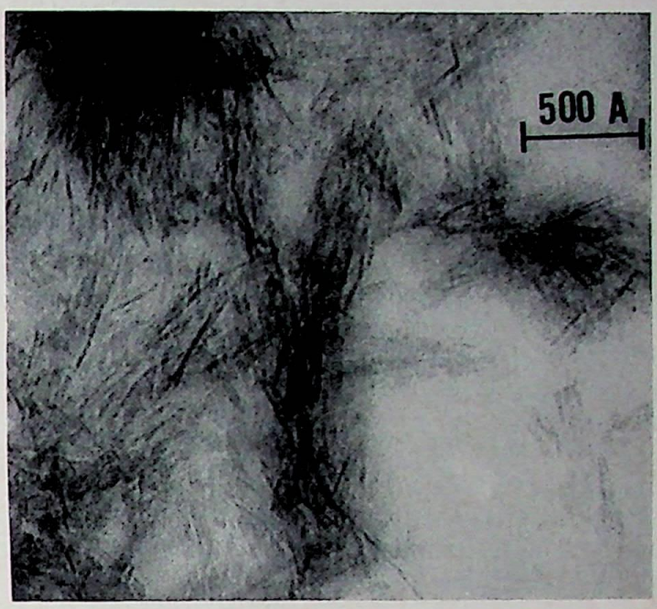


FIG. 33. Higher magnification of an area in a sample similar to Fig. 31, where the rod-like crystals predominate. ×260 000.

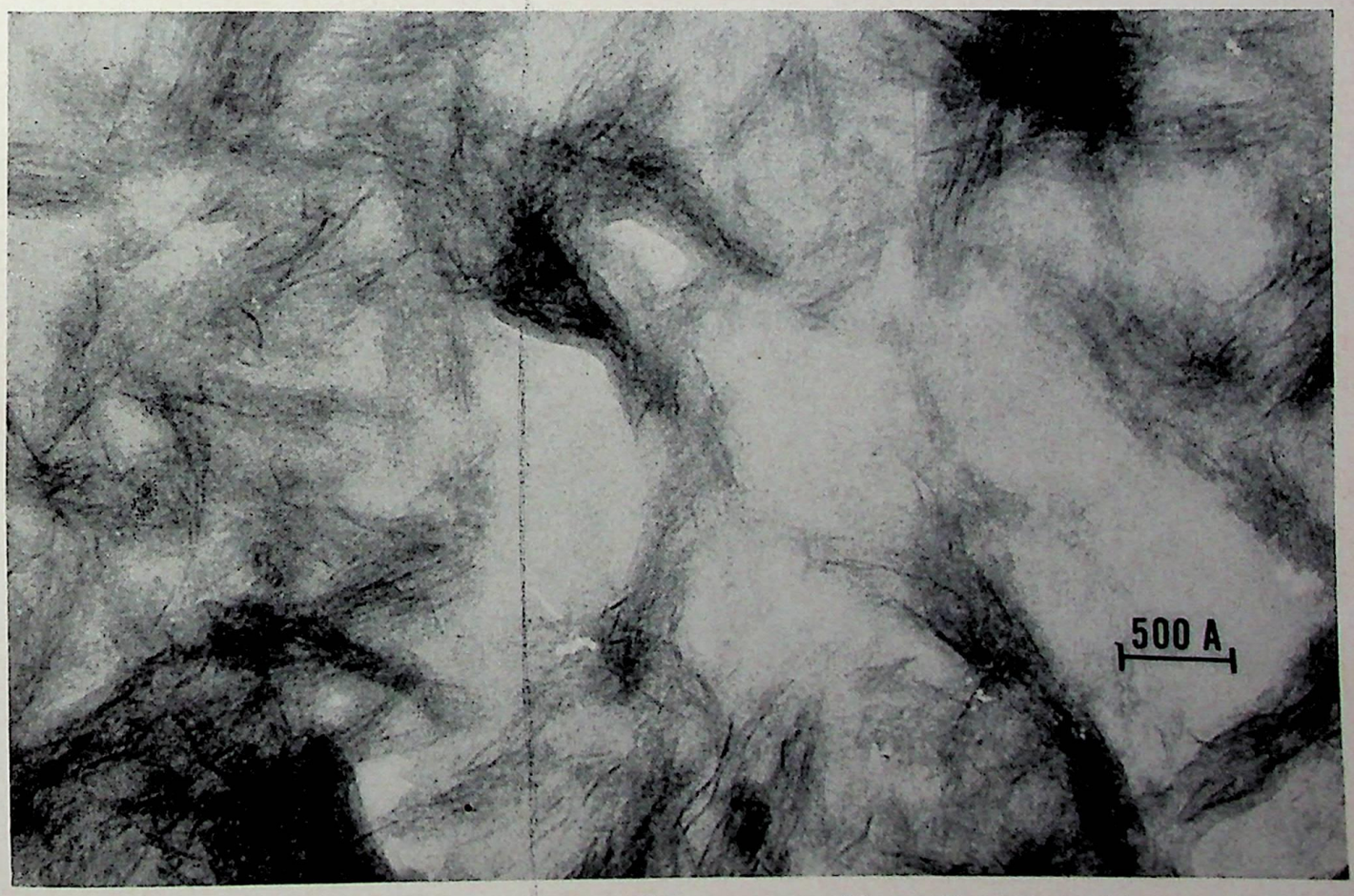


FIG. 34. *In vitro* precipitated apatite crystals. Note the rod-like appearance without evidence of plate formation, except for the minimal background electron density. Crystals vary from 8 to 20 A in width, and many appear to be "bent" similar to areas of lightly calcified bone. ×260 000.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

the fibril. These crystals were, therefore, in no way capable of being oriented by *the individual fibrils*.

In samples of reconstituted collagens, the interfibrillar spaces were large, and the over-all organization of the fibrils was also not able to influence the orientation of the crystals (similar to cartilage). In the case of demineralized bone, however, the collagen fibrils are more closely packed and well oriented. Under these circumstances, the elongated plate-like crystals mechanically orient with their flat, thin faces between adjacent fibrils and their long dimensions approximately parallel to the collagen fiber axis. This arrangement gives a co-oriented x-ray diffraction pattern of apatite and collagen but not as good a one as that in native bone, where the majority of the crystals are within the fibrils.

Experiments are now underway in which the physicochemical conditions during crystal *growth* as well as crystal *induction* are being carefully controlled in an attempt to grow the crystals primarily within the collagen fibrils similar to that of native bone.

Crystal Nonstoichiometry

As mentioned in the beginning of this paper, one of the perplexing problems in the study of the nature of the bone crystals or of *in vitro* precipitated apatites has been their nonstoichiometry. Recall that, in bone and in some *in vitro* precipitated crystals, the average width of the rods or of the stacked plates (or both) varied from 10 to 30 A. This dimension corresponds to the *a*- or *b*-axis of the unit cell which is approximately 9.43 A. Thus, the crystals are composed of 1 to 3 unit cells in this dimension. Although the unit cell of hydroxyapatite may be represented as in Fig. 1, it must be remembered that it depicts only *conceptually* the *spatial relations* of the constituent atoms and molecules, and that many of the individual atoms and molecules are shared by adjoining unit cells in an actual crystal.

Therefore, the stoichiometry of the solid phase may be represented by the structure of the unit cell *only* when there is a very large number of unit cells comprising the crystal, so that the unit cells or the portions of them making up the surface are not statistically significant. When the entire crystal is composed (in any dimension or dimensions) of only a few unit cells, however, and if the planes on which crystal growth ceases are relatively uniform along the length of the crystal, *it is not possible to obtain the theoretical stoichiometry*. The *actual* stoichiometry in these cases will be determined by the atomic planes in which crystal growth ceases. If the crystal surface planes on which crystal growth ceases vary a good deal, it is possible that a fortuitous combination might occur which would result in a theoretical stoichiometry, but this is highly unlikely. In *in vitro* precipitated crystals, however, there is ordinarily a very wide range of crystal size and shape (and probably hydration), and the statistical summation of the calcium and phosphorus content of such an heterogeneous sample, whether stoichiometric or not, would not be very meaningful. Although there is obviously a number of other previously mentioned factors such as internal lattice substitution, surface adsorption, vacant lattice sites, etc., which may influence the stoichiometry, the point is that, *if none of these existed, it still would not be possible to obtain stoichiometric hydroxyapatite* with crystals the size of those obtained in bone or prepared *in vitro*.

The larger "defect" apatite crystals reported by Posner and Perloff[11] of course poses another problem, but it is possible that such large crystals represent a somewhat different structure than hydroxyapatite and are similar to OCP.

The small size of the crystals also may be an important factor in determining the amount of water associated with apatite crystals precipitated from aqueous solutions as discussed on p. 361. This "excess" water is many times greater than that adsorbed by initially dry crystals from a vapor phase of water and consists of many more molecular layers of water than that usually considered possible for a surface to bind because of electrical-field effects.

In addition to the fact that some of the water may enter the lattice structure itself, since the minute apatite crystals in aqueous suspension tend to stick together and form stacks and bundles, the very large surface tension and capillarity effects between such crystal surfaces could well "trap" a large amount of water (in addition to a bound monolayer) and still resist separation from the crystals by mechanical centrifugation.

Closing Comments

As regards other mineralized biological tissues, one can see from Table I that the various crystals are anatomically related to specific organic matrices. Note the wide variety both of inorganic crystals and of organic matrices in which mineralization occurs. In fact, in some marine mollusks, calcium carbonate exists in two of its three possible crystallographic forms (calcite and aragonite) *in the same shell*, in well-demarcated regions which *border* on each other.

Another example of such an intimate chemical and ultrastructural relationship between the inorganic and organic phases is that which occurs in oyster shells. Electron micrographs have shown a close and ordered relationship between the calcium carbonate crystals and the organic matrix of conchiolin.[88–90] Here too, calcification is initiated by small seed-like crystals which are initially deposited in a regular fashion in the organic matrix and appear similar to synthetically grown crystals from *supersaturated* solutions.[90]

Although the mechanisms by which mineralization is initiated and regulated in the organism are undoubtedly far more complex than those in the relatively simple model system described, the process of the heterogeneous nucleation of inorganic crystals by highly specific regions in the organic matrix as the result of a characteristic stereochemical array of certain reactive groups is

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

probably a fundamental one, not only for the calcification of bone and cartilage but also for biological mineralization in general.

APPENDIX A. SOME THERMODYNAMIC AND KINETIC CONSIDERATIONS IN PHASE TRANSFORMATIONS‖

Crystallization is a specific case of a more general phenomenon in which a change in state involves a phase transformation. A complete consideration of phase transformations would encompass (1) the conditions under which a system can or cannot persist in a distinct, physically homogeneous state (a single phase); (2) a *mechanism* for the formation of new phases from such systems; and (3) a quantitative description of the time-dependence of the process.

Since the primary purpose of this exposition, however, is to provide a background of some basic concepts in phase transformations in order to understand better the general phenomenon of crystallization and of biological mineralization, attention is directed primarily to some general considerations of the first two aspects.

Equilibrium and the Stability of Phases¶

In most general terms, Gibbs[91] has defined equilibrium as a state independent of time, that is, a state in which *all* of the sensible properties which describe the system do not vary with time. Thus, in formulating both the necessary and sufficient conditions of equilibrium in terms of thermodynamic functions (e.g., energy, entropy, free energy, etc.), Gibbs considered equilibrium with respect to *all possible variations* of the state of the system, and *not* with respect to *certain* variations only. In practice, however, it is necessary to idealize most systems by imposing certain conditions of restraint and then by studying them with respect to *certain* possible variations only. In the case at hand, the possible variation of interest is the formation of entirely new phases from an initially homogeneous phase.

The usual criteria of equilibrium [for example, that for all possible variations in the state of an isolated system $(\delta E)_S \geqq 0, (\delta S)_E \leqq 0, (\delta F)_{p,T} \geqq 0$, etc.] do not give the information necessary for the interpretation of phase transformations, since many systems which meet these general criteria of equilibrium may vary widely in their *relative* tendency or ability to form new phases. On the basis of this *relative stability*, Gibbs[91] distinguished four kinds of equilibrium: (1) stable, (2) neutral, (3) unstable, and (4) metastable equilibria.

‖ The reader is referred to the following article for a more comprehensive review of the subject: D. Turnbull, Advances in Solid-State Phys. 3, 225 (1956).

¶ The material presented here has been compiled from a series of lecture notes from a graduate course in Chemical Thermodynamics, based on the writings of J. Willard Gibbs, given at the Massachusetts Institute of Technology, Department of Physical Chemistry, by Professor James A. Beattie. I would like to thank Professor Beattie for permission to use some of this material, including Figs. 35 and 36, and for many helpful suggestions and criticisms in the preparation of the Appendix.

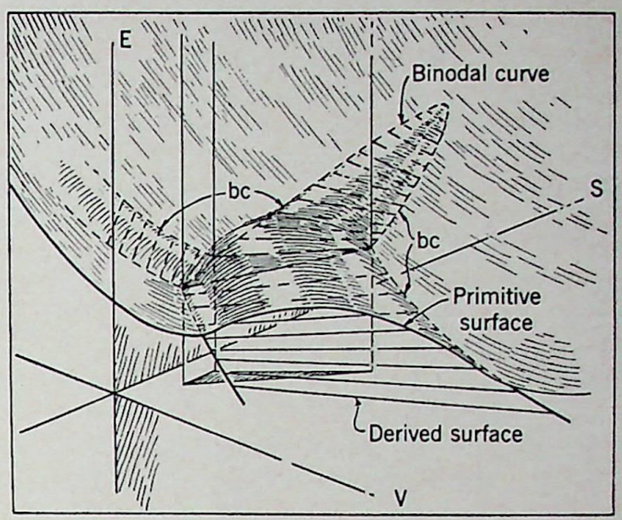


FIG. 35. Artist's (P. Lund) conception of the Gibbs energy-entropy-volume surface for a one-component system. The dotted lines on the primitive surface connect points of common tangency (neutral equilibrium), and the construction of a derived surface from such points is illustrated also. The orthogonal projection of the plane triangle on the S-V plane, representing the triple point, is shown. bc represents binodal curves.

The fundamental principles can, with respect to the stability of phases, conveniently and easily be visualized by reviewing certain pertinent features of the Gibbs energy-entropy-volume thermodynamic surface. For purposes of exposition, consider the special case of a pure substance when effects produced by gravity, electricity, capillary tensions or distortion of solid phases are absent, or may be neglected.

For all possible infinitesimal changes of state involving only expansion work in a closed system, the first and second laws of thermodynamics state that

$$dE = TdS - pdV. \tag{A1}$$

Integration of (1) gives the equation

$$E = E(S,V). \tag{A2}$$

Thus, the relations between the energy, entropy, and volume may be represented by a surface in space whose rectangular coordinates are represented by the volume, energy, and entropy of the system (Fig. 35). This surface, called the *primitive surface*, represents equilibrium states of a homogeneous phase. At any point on the primitive surface, the equilibrium temperature and pressure are defined by the slopes of the surface at that particular point:

$$\left(\frac{\partial E}{\partial S}\right)_V = T, \left(\frac{\partial E}{\partial V}\right)_S = -p \tag{A3}$$

and the tangent plane at the point given by

$$E = E_0 + T_0(S - S_0) - p_0(V - V_0). \tag{A4}$$

The intersection of the tangent plane with various other planes and the coordinate axes give various thermodynamic functions. For example, its intersection with the E axes gives F_0, the Gibbs free energy of the substance at the point P_0.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

States of the system consisting of two or more phases in equilibrium (heterogeneous equilibrium) can also be represented in E-S-V space, since the system as a whole has a definite energy, entropy, volume, free energy, etc., as well as temperature and pressure. In heterogeneous equilibrium, the substance in each of its aggregation states has the same temperature, pressure, and molal free energy. Thus, planes tangent to points on the primitive surface representing, for example, one mole of the substance in each of these aggregation states will have the same slope $(-p,T)$ and the same intercept in the E-axes (free energy). Hence, these points have a common tangent plane.

In the case of two phases (say liquid and vapor) in equilibrium at a definite temperature and pressure, there are two points on the primitive surface, one representing one mole of liquid and the other representing one mole of vapor at this pressure and temperature which have a common tangent plane. Lying along the straight line connecting the points are the states of mixtures of liquid and vapor in equilibrium at this particular temperature and pressure. By taking all of the lines determined by such tangent planes for varying temperatures and pressures, another surface may be generated called the *derived surface* which represents states of heterogeneous equilibrium. One may conceive of this derived surface as being produced by rolling a double tangent plane (tangent to the primitive surface at two points) on the primitive surface and by connecting the successive pairs of *conjugate points* of tangency by a series of straight lines (Fig. 35).

In the case of three phases in equilibrium (i.e., liquid, vapor, solid), there are three points on the primitive surface with a common tangent plane, and states of the substance at the triple point (mixtures of all three phases in equilibrium) are presented in E-S-V space by points in a plane triangle (Fig. 35). The derived surfaces, therefore, include all of the plane triangles representing three phases in equilibrium, and all of the developable surfaces representing two phases in equilibrium. It is a continuous ruled surface (generated by the motion of a straight line), but it is *not* cylindrical.

Properties of the Thermodynamic Surfaces Which Indicate the Stability of Thermodynamics Equilibrium

Consider a specified amount of a pure substance A, immersed in a large medium M, at constant pressure p_0 and temperature T_0 with an initial energy, entropy, and volume (E',S',V') and which undergoes the following change of state

$$A(E',S',V')=A(E'',S'',V'')$$

under conditions where the action of the system on the media is substantially reversible. It can easily be shown from the first and second laws of thermodynamics that the following relation must hold:

$$(E''-T_0S''+p_0V'')\leqq(E'-T_0S'+p_0V'). \quad \text{(A5)}$$

With reference to the thermodynamic surface, the terms enclosed in parentheses are the vertical distances of the points (E'',S'',V'') and (E',S',V') representing the final and initial states of the system above a plane passing through the origin having the slopes $-p_0,T_0$. Geometrically, it defines the conditions under which changes in state can occur spontaneously: where this vertical distance representing the final state is *less* or at most *equal* to the distance representing the initial state. Changes in state which result in an increase of this distance are, therefore, *not* possible. This principle is crucial to an understanding of phase changes, for it clearly delineates the regions in E-S-V space (and, therefore, the equilibrium of such states) where phase transitions are possible.

In examining the possible equilibrium states of a pure substance in a medium at temperature T_0 and pressure p_0 (represented by points on the primitive surface), two possible categories of change must be considered. The behavior of systems may be related to (1) *continuous changes*—that is, changes considered with respect to nearby or adjacent states, and (2) *discontinuous changes*—that is, in relation to states at a finite distance from the point in question. The *curvatures* of the primitive surface determine the former, while the *over-all relation* of the tangent plane to the surface determines the latter.

Stable Equilibrium

If the primitive surface falls *above* the fixed tangent plane, *except* at the single point of contact (representing the initial state), the state of the system represented by that point is "absolutely stable" toward a phase change when in a medium of constant pressure and temperature. That is, there are no points in E-S-V space for which Eq. (5) is satisfied, and there are, therefore, no possible changes in state which can occur. Any "unnatural" perturbations of the system (such as, a local fluctuation in density) would necessarily lead to changes in state represented by points in E-S-V space where the distance of any such point above the tangent plane would be greater than that of the original point. Since Eq. (5) has shown that such changes in state are *not* possible, natural processes will occur which will return

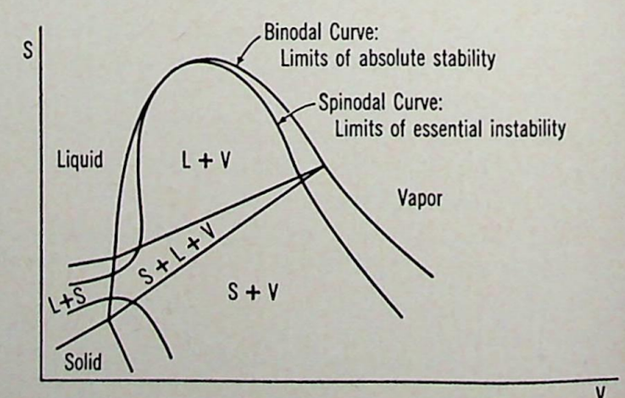


FIG. 36. Orthogonal projection of the limits of absolute stability (binodal curve) and limits of essential instability (spinodal curve) on the S-V plane for a substance having one solid phase.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

the system to its original state. It can be shown that that part of the primitive surface representing such stable states of equilibrium is concave upward in both of its principal curvatures.

The points of tangency of the rolling double-tangent plane representing two phases in equilibrium trace the *binodal curves* (Fig. 36) or the limits of *absolute stability*. The tangent plane to points on the primitive surface which fall *outside the binodal curves* has the primitive surface entirely above it except at the single point of contact. That portion of the primitive surface outside the binodal curves represents states of *absolute stability*. This part of the primitive surface (referred to as the *surface of absolute stability*) together with the derived surfaces constitute the *surface of dissipated energy*. The tangent plane to a point on the surface of *absolute stability* is always below the surface of dissipated energy except at the point of tangency.

Solutions in such states of equilibria would be absolutely stable with regard to the *formation* of new phases. The mechanical analog of such a system may be represented by a marble in a hemispherical bowl fitted with a cover, with the conditions of restraint that the marble cannot sink through the bowl or be otherwise removed from it [Fig. 37(a)].

Neutral Equilibrium

If the primitive surface does not fall anywhere below the fixed tangent plane but meets it at more than one point, the equilibrium of such states of the system is considered *neutral*. In such cases, if the system in such an equilibrium state is changed from its original state to a new state represented by another point of tangency to the primitive surface, the distances of both points above the fixed tangent plane (at constant T and p) obviously will be equal (zero). Therefore, such systems will have no tendency to pass into one of the other states (as represented by other points of common tangency) or to return to their original state if so displaced. However, such systems are stable with respect to continuous changes in state similar to systems in stable equilibrium.

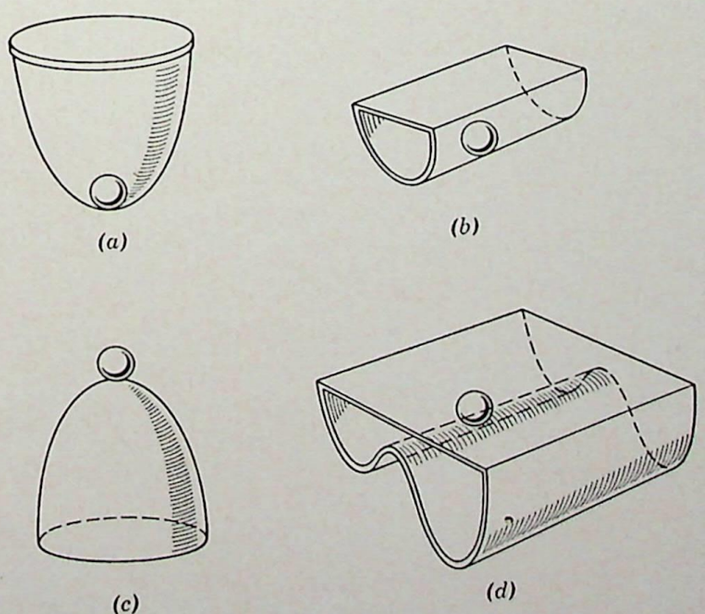


FIG. 37. Diagrammatic illustration of the four types of equilibria based on their relative stability.

A marble in a horizontal trough is an example of such a state of neutral equilibrium [Fig. 37(b)]. Although the displacement of the marble up the sides of the container eventually results in the return of the marble to its original position at the bottom of the trough, there are a number of positions of the marble along the bottom of the trough where no spontaneous tendency to change exists and where no spontaneous tendency for a return of the marble occurs if such horizontal displacements are made. Solutions in such states of neutral equilibrium are, therefore, also stable with regard to the *formation* of new phases.

Unstable Equilibrium

If the primitive surface be continuous, there must necessarily exist regions between the binodal curves where the curvature of the primitive surface is concave *downward* in at least one of its principal curvatures. Points *on* the binodal curves (representing states of neutral equilibrium) separate such states of stability and instability with respect to *discontinuous* changes. The lines on the primitive surface dividing the portion which is concave *upward* in both of its principal curvatures from the portion which is concave downward in one or both of its principal curvatures represent the limit of *essential instability* or the *spinodal curves* (Fig. 36).

In such regions, it is obvious that, where part of the surface falls below the fixed tangent plane, it is possible to change the initial state of the system such that the point representing the final state is now below the fixed tangent plane. In this case, according to Eq. (5) natural processes occur which cause the system to continue to change (represented by moving the point further from the tangent plane) until a state is reached which is entirely different from the initial state [Fig. 37(d)]. Such states of equilibrium are *unstable* with respect *both* to continuous and discontinuous changes. That part of the primitive surface which lies inside the spinodal curve represents such states of *unstable* equilibrium. This type of equilibrium is rarely—if ever—attained in practice, and solutions in such a precarious state of equilibrium would hardly remain so for very long periods.

Metastable Equilibrium

The primitive surface which lies *between* the spinodal and binodal curves is concave upward in both of its principal curvatures and is *stable* with respect to *continuous changes* (adjacent states). However, it is so situated with respect to the over-all thermodynamic surface that other states do exist a finite distance away, where the primitive surface falls *below* the tangent planes drawn through points on the primitive surface

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

which lie between the binodal and spinodal curves. Such systems which are *stable* with respect to *continuous changes* (adjacent states) and *unstable* with respect to *discontinuous* changes (distant states) are said to be in *metastable equilibrium*. If the conditions necessary for such a discontinuous change *were not* present, the equilibrium would remain stable *indefinitely*. But, for example, if small portions of the same substance in one of the more stable states of aggregation, represented by points below the tangent plane, are introduced or otherwise caused to form by very small disturbances (perhaps ones that cannot be detected experimentally), the equilibrium would be destroyed and a change of state (a phase change) would occur [Fig. 37(d)]. Solutions in metastable equilibrium, therefore, although *capable* of remaining stable indefinitely, *can* form new, more-stable phases under certain conditions without changing the entire system.

With reference to the thermodynamic surface, it is thus apparent, since points *on* the surface of dissipated energy represent *stable* or *neutral* equilibrium and since points *below* the surface of dissipated energy have no physical significance, that points on the primitive surface which lie *above* the surface of dissipated energy representing states of *unstable* or metastable equilibrium delineate the regions in E-S-V space from which phase changes can occur at constant temperature and pressure.

On the basis of these conditions, it can be shown that a single analytical expression representing both the necessary and sufficient conditions of equilibrium for a system in a medium of temperature T_0 and pressure p_0 is

$$\delta(E-T_0S+p_0V)\equiv\delta F=0 \qquad \text{(A6)}$$

for all variations of the state of the system. The system is absolutely *stable* with respect to first-order changes if δF is a minimum, and it is *unstable* if δF is a maximum. Finally, if Eq. (6) holds, but for *some finite changes*, the inequality $\Delta F<0$ exists, the state is *metastable*.

Although the foregoing discussion is limited to systems of one component, Gibbs (reference 91, pp. 100–115) has extended his treatment to a study of the internal stability of homogeneous fluids of many components.

Mechanisms and Time Dependence of Phase Transformations

The discontinuous change of metastable or unstable systems resulting in the formation of the initial fragments of a new, more stable phase is called *nucleation*. When such a phase-change occurs in the interior of a metastable or unstable system in the absence of structural impurities, it is referred to as *homogeneous nucleation*, whereas a phase-change initiated *by* and *on* foreign inclusions extraneous to the system is called *heterogeneous nucleation*.

Homogeneous Nucleation

Classical theories of homogeneous nucleation have assumed that phase-changes occur by the formation of intermediaries as a result of local fluctuations of certain properties (such as density) in initially homogeneous phases in metastable or unstable equilibrium.[46,92–94]

These intermediaries consisting of clusters of molecules (or ions) of the initial phase vary in size (and possibly composition, shape, and structure). They are referred to as "embryos" and are considered to exist as a true heterogeneous system with the mother phase. The cluster of critical size, composition, structure, etc., capable of *further* spontaneous growth (with a net decrease in free energy) and, therefore, capable of initiating the formation of the new phase, is called a *nucleus* (Fig. 38).

The embryos and nuclei are considered to arise by the stepwise addition of single molecules, i.e., a *bimolecular mechanism*.[46] The *order* of the reaction, therefore, is equal to the number of molecules in the nucleus.

As first proposed by Volmer and Weber,[93] the theory considered that nuclei can exist in a state of unstable equilibrium with the mother phase, making it possible, at least in principle, to define the nucleus in terms of reversible thermodynamics. Considering surface effects, the free energies of formation of spherical clusters can be calculated as a function of the radius (Fig. 38). It is obvious from Fig. 38 that the over-all free energy goes through a maximum which corresponds to the critical size cluster (nucleus) of radius, r^*, which can grow spontaneously with a net decrease in free energy. From fluctuation theory, the probability of nuclei formation and, therefore, also of their concentration can be approximated since they are proportional to $\exp(-\Delta F^*/kT)$, where ΔF^* is equal to the free energy or the work of formation of a nucleus corresponding to the maximum as shown in Fig. 38. The condensation velocity or nucleation rate is then evaluated by computing the collision frequency of single molecules with the nuclei. This free-energy barrier is, therefore, similar to the activation energy of ordinary chemical reactions in permitting the derivation of a *nucleation rate*.††

Becker and Döring[45] treated the problem differently. They assumed a size distribution of "embryos" and evaluated the rate of condensation by the solution of a set of equations relating how the number of embryos of any particular size changed with time. By assuming a steady state, the mechanism of change was, therefore, simply the kinetic process of unit condensation and unit evaporation.

In both of the cases, expressions for nucleation rate were quite similar [$J\sim\exp(-1/\ln S^2)$], where J is the nucleation rate and S the supersaturation ratio p/p_e where p is the actual vapor pressure and p_e the equilibrium vapor pressure), and indicated a very marked

†† In more-complicated systems, such as occurs in the case of inorganic crystallization, the various-sized embryos may have a number of compositional and structural differences which also involve free-energy changes, and, in particular instances, there may be important effects owing to strain energy as well. Thus, over-all free-energy barrier to nucleation will be provided by a combination of the chemical-, interfacial-, and strain-free energies.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

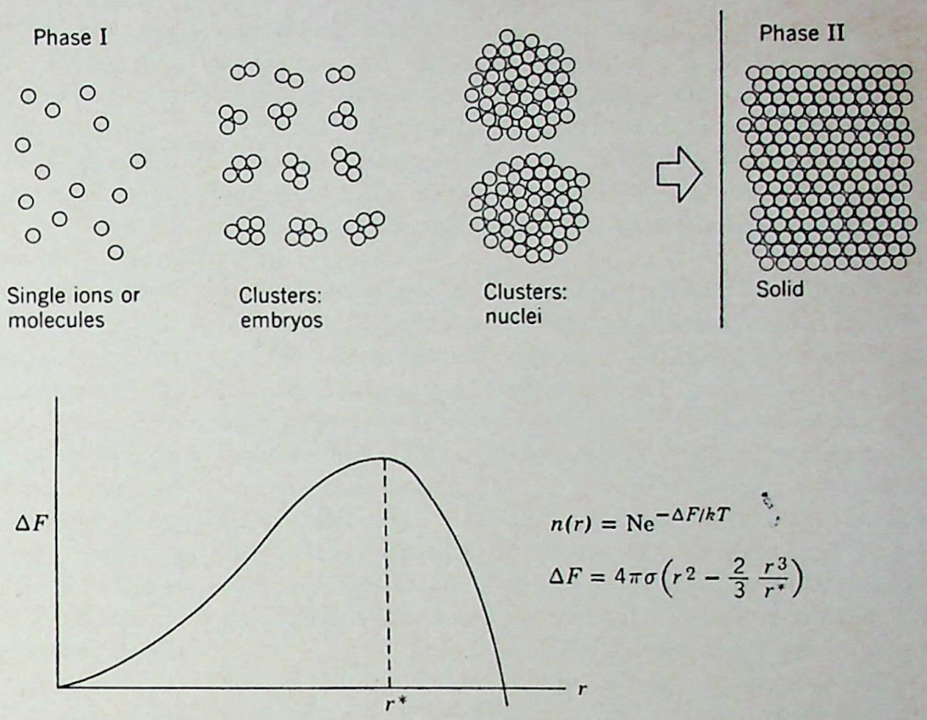


FIG. 38. Schematic illustration indicating mechanism of nucleation as proposed principally by Volmer *et al.*[46,93,94] and by Frenkel.[92]**

dependence of nucleation rate upon the degree of supersaturation, particularly near the critical supersaturation ratio (i.e., where the nucleation rate rather suddenly becomes sufficiently large for ready measurement). For example, in the case of water vapor condensing to liquid drops, the time that must elapse for the appearance of the first drop at a supersaturation ratio of 4, is 0.1 sec, whereas increasing the supersaturation ratio to 5 decreases the time to 10^{-13} sec, and decreasing the ratio to 3 increases the time to 10^3 years![46]

Although based on the theory of Volmer and Weber (fluctuation theory), it is possible that homogeneous nucleation can occur at any level of metastability. Certain theoretical considerations (as well as experimental findings) indicate that this is not possible, except at the limit of essential instability or from systems in *unstable* equilibrium rather than in *metastable* equilibrium. The difficulties arise conceptually from the definitions of equilibrium (a state independent of time), of metastability (stable with respect to continuous changes in state, but unstable with respect to certain discontinuous changes in state), and of what constitutes a discontinuous change in state, and from the applicability of $e^{-\Delta S/k}$ to predict the probability of fluctuations which are large enough to be considered a new phase. Thus, Frenkel[92] considers that the density fluctuations which lead to the formation of embryos (heterophase) transcend the limits usually considered in the ordinary statistical theory of homogeneous systems wherein ordinary fluctuations (homophase) lie "within the limits compatible with the preservation of a given aggregation state" (Frenkel,[92] p. 375).

** While this is usually considered the mechanism of nucleation for most systems, there is recent evidence that, in the nucleation of solid-state transformations[95] the nucleii arise by a cooperative phenomenon and not by discrete atomic jumps.

From these considerations, it would appear that nucleation from true metastable systems occurs only as a result of the introduction of heterogeneities and not spontaneously. This would also seem to be Gibbs' interpretation of the stability of metastable systems, since he states:

". . . the mass in question must be regarded as in strictness stable with respect to the growth of a globule of the kind considered, since W, the work required for the formation of such a globule of a certain size (viz, that which would be in equilibrium with the surrounding mass), will always be positive. Nor can smaller globules be formed, for they can neither be in equilibrium with the surrounding mass, being too small, nor grow to the size of that to which W relates. If, however, by *any external agency* [Ital.: Ed.] such a globular mass (of the size necessary for equilibrium) were formed, the the equilibrium has already (page 243) been shown to be unstable, and with the least excess in size, the interior mass would tend to increase without limit except that depending on the magnitude of the exterior mass" (Gibbs,[91] pp. 255–256).

Heterogeneous Nucleation

The discontinuous change in state which results in nuclei formation by the introduction of a foreign inclusion may be the result of a number of different effects. Thus, either by virtue of nonspecific surface forces or by very specific interactions between certain groups on the surface and the molecules or ions of the initial phase, such heterogeneities may adsorb or bind the molecules or ions of the initial phase on their surface; by acting either as a core whose surface is now composed of the molecules of the initial phase or by binding the molecules or ions in a particular steric array, they may form

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

a cluster of critical size, shape, and configuration necessary for nucleation. Since the critical-size cluster (assuming no structural or compositional changes) varies with the supersaturation ratio, nucleation rates for heterogeneous nucleation will, as in the case of homogeneous nucleation, be strongly dependent upon and markedly dominated by the degree of metastability of the system.

With regard to crystallization, it is of interest to note the findings of Turnbull and Vonnegut[63] who have proposed a crystallographic theory of crystal nucleation. The theory emphasizes the importance of *geometric* and *structural* factors of nucleation catalysts with regard to their catalytic potency and ability to initiate phase-changes. Based on a number of their own observations as well as on experiments of others, they have postulated that those substances acting as effective nucleation catalysts have atomic arrangements and lattice spacings on certain low-index planes which are very similar to those of the crystal being nucleated.

The potency of the catalyst is formulated to be directly proportional to the reciprocal of the disregistry between the lattice parameters of the catalytic surface and the forming crystal on these planes. Thus, in the formation of ice crystals from supersaturated droplets of water vapor for example, the most potent nucleation catalyst was found to be AgI[96,97] whose lattice structure and atomic arrangement are remarkably similar to ice. Similarly, in the case of the nucleation of hydrated sodium-sulfate ($Na_2SO_4 \cdot 10H_2O$) by sodium-tetraborate ($Na_2B_4O_7 \cdot 10H_2O$) crystals, it was noted that both crystals belong to the same space group and that the disregistry was very small between the two types of crystals on certain basal planes.[98] That some heterogeneities will not act as catalysts at all is probably related to such specific interactions and geometrical factors.

Although many considerations such as specific interactions, types of chemical bonds, etc., as well as relatively nonspecific factors such as surface dislocations and defects, influence the potency or ability of the catalyst in any particular case, the importance of *geometric* and *structural* factors should be kept in mind because of their possible implications for the theory of biological mineralization which is presented.

Other Related Considerations

Whereas *nucleation*—referring to the process of *forming* the infinitesimally small fragments of the new phase—can *in concept* be separated from the subsequent *growth* of these fragments into macroscopic particles of the new phase, *nucleation rate*—which is usually determined by measuring the appearance of the macroscopic particles of the new phase—*operationally* is not completely independent of the kinetics of such growth.

Therefore, the *operational* definition of *nucleation rate* is not completely consistent with the *theoretical* concept of nucleation rate, since the latter strictly speaking refers only to the rate at which the nuclei are forming and not to the subsequent growth of the nuclei to crystals, nor to the phenomenon of recrystallization.

Furthermore, measurements of nucleation rate, particularly in the case of heterogeneous nucleation have given little or no evidence of the molecular or atomic sequence of the nucleation phenomena itself—i.e., the mechanism of nuclei formation—which is of particular importance in the case of biological mineralization, where certain of these steps may be enzymatically controlled and regulated, for example.

Moreover, in the case of crystallization, the number and size of crystals is determined by the relative rates of crystal growth and nucleation rate once nucleation has been initiated. Considered independently, crystal growth may vary in just the opposite fashion as does nucleation rate with respect to certain variables such as temperature. In addition, the growth of crystal nuclei is dependent not only on diffusion rate (i.e., supersaturation ratio) but also on the presence of surface defects (e.g., growth by screw dislocations) even at low-supersaturation ratios.[99]

An interpretation of the entire phenomenon of the crystallization of a specific mineral would, therefore, require not only a study of nucleation phenomena but also considerations of the factors influencing crystal growth, crystal size, crystal habit, etc., for a particular crystal structure. Furthermore, the conditions influencing *recrystallization* (the growth of large crystals at the expense of smaller ones), might also be quite important in the over-all process even after the solid state had been achieved.

BIBLIOGRAPHY

[1] J. E. Eastoe and B. Eastoe, Biochem. J. **57**, 453 (1954).

[2] J. E. Eastoe in *The Biochemistry and Physiology of Bone*, G. H. Bourne, editor (Academic Press, Inc., New York, 1956), p. 81.

[3] Author's unpublished observations.

[4] W. F. DeJong, Rec. trav. Chim. **45**, 445 (1926).

[5] D. Carlström, Acta Radiol. Suppl. 121 (1955).

[6] W. F. Neuman and M. W. Neuman, Chem. Revs. **53**, 1 (1953).

[7] P. Cartier, Bull. soc. chim. biol. **30**, 65 (1948).

[8] M. J. Dallemagne and H. Brasseur, Acta Biol. Belgica **2**, 440, 445 (1952).

[9] H. Brasseur, M. J. Dallemagne, and J. Mélon, Bull. soc chim. France **16**, 109 (1949).

[10] M. J. Dallemagne, C. Fabry, and A. S. Posner, J. physiol. (Paris) **46**, 325 (1954).

[11] A. S. Posner and A. Perloff, J. Research Natl. Bur. Standards **58**, 279 (1957).

[12] D. Carlström, Biochim. et Biophys. Acta **17**, 603 (1955).

[13] S. B. Hendricks and W. L. Hill, Proc. Natl. Acad. Sci. U. S. **36**, 731 (1950).

[14] J. H. Weikel, Jr., and W. F. Neuman, U. S. Atomic Energy Report, University of Rochester, UR-228 (1952).

[15] R. Kunin, K. L. Elmore, and D. L. Johnson, TVA Report, unpublished. Data reported by W. F. Neuman and M. W. Neuman in *The Chemical Dynamics of Bone Mineral* (The University of Chicago Press, Chicago, 1958), p. 39.

[16] W. F. Neuman and M. W. Neuman, *The Chemical Dynamics of the Bone Mineral* (The University of Chicago Press, Chicago, 1958).

[17] M. J. Dallemagne and J. Mélon, Arch. de Biol. **57**, 79 (1946).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

[18] P. W. Arnold, Trans. Faraday Soc. **46**, 1061 (1950).
[19] W. E. Brown, J. R. Lehr, J. P. Smith, and A. W. Frazier, J. Am. Chem. Soc. **79**, 5318 (1957).
[20] A. S. Posner (personal communication, 1958).
[21] J. W. Gruner and D. McConnell, Z. Kristallog. **97**, 208 (1937).
[22] D. McConnell in *Metabolic Interrelations*, Transactions 4th Conference, E. C. Reifenstein, Jr., editor (Josiah Macy, Jr., Foundation, New York, 1952).
[23] L. Huber and C. Rouiller, Experientia **7**, 338 (1951).
[24] C. Rouiller, L. Huber, E. Kellenberger, and E. Rutishauser, Acta Anat. **14**, 9 (1952).
[25] W. J. Schmidt, Handbuch biol. Arbeitsmethoden, **A5**, T10, 435 (1934).
[26] J. Wolff, *Das Gesetz der Transformation der Knochen* (Quarto, Berlin, 1892).
[27] R. A. Robinson, J. Bone and Joint Surg. **34A**, 389 (1952).
[28] R. A. Robinson and M. L. Watson, Anat. Record **114**, 383 (1952).
[29] R. A. Robinson and M. L. Watson, Ann. New York Acad. Sci. **60**, 596 (1955).
[30] S. Fitton-Jackson, Proc. Roy. Soc. (London) **B146**, 270 (1957).
[31] H. Sheldon and R. A. Robinson, J. Biophys. Biochem. Cytol. **3**, 1011 (1957).
[32] R. Robinson, *The Significance of Phosphoric Esters in Metabolism* (New York University Press, New York, 1932).
[33] E. Freudenberg and P. Györgi, Biochem. Z. **118**, 50 (1921).
[34] A. B. Gutman and T. F. Yü, in *Metabolic Interrelations*, Transactions 1st Conference (Josiah Macy, Jr., Foundation, New York, 1950), p. 11.
[35] A. E. Sobel and M. Burger, Proc. Soc. Exptl. Biol. Med. **87**, 7 (1954).
[36] A. E. Sobel, Ann. N. Y. Acad. Sci. **60**, 713 (1955).
[37] P. S. Rubin and J. E. Howard in *Metabolic Interrelations*, Transactions 2nd Conference (Josiah Macy, Jr., Foundation, New York, 1951), p. 155.
[38] B. S. Strates, W. F. Neuman, and G. J. Levinskas, J. Phys. Chem. **61**, 279 (1957).
[39] B. S. Strates and W. F. Neuman, Proc. Soc. Exptl. Biol. Med. **97**, 688 (1958).
[40] M. J. Glimcher, A. J. Hodge, and F. O. Schmitt, Proc. Natl. Acad. Sci. U. S. **43**, 860 (1957).
[41] M. B. Engel, N. R. Joseph, and H. R. Catchpole in *Metabolic Interrelations*, Transactions 5th Conference, E. C. Reifenstein, Jr., editor (Josiah Macy, Jr., Foundation, New York, 1953), p. 105.
[42] I. Gersh in *Connective Tissues*, Transactions 2nd Conference (Josiah Macy, Jr., Foundation, New York, 1952), p. 11.
[43] F. McLean and M. R. Urist, *Bone, An Introduction to the Physiology of Skeletal Tissue* (The University of Chicago Press, Chicago, 1955).
[44] F. McLean in *Metabolic Interrelations*, Transactions 2nd Conference (Josiah Macy, Jr., Foundation, New York, 1950), discussion p. 101.
[45] R. Becker and W. Döring, Ann. Physik **24**, 719 (1935).
[46] M. Volmer, *Kinetik der Phasenbildung* (Edwards Brothers, Ann Arbor, 1945).
[47] W. F. Neuman in *Metabolic Interrelations*, Transactions 2nd Conference (Josiah Macy, Jr., Foundation, New York, 1950), p. 32.
[48] Author's unpublished results.
[49] D. S. Jackson (personal correspondence, 1958).
[50] C. C. Solomons and J. T. Irving, Biochem. J. **68**, 499 (1958).
[51] B. F. Martin and F. Jacoby, J. Anat. **83**, 351 (1949).
[52] P. Doty and T. Nishihara in *Recent Advances in Gelatin and Glue Research*, G. Stainsby, editor (Pergamon Press, London, 1958), p. 92.
[53] P. J. Flory and R. R. Garrett, J. Am. Chem. Soc. **80**, 4836 (1958).
[54] W. Cochran, F. H. C. Crick, and V. Vand, Acta Cryst. **5**, 581 (1952).
[55] F. H. C. Crick, Acta Cryst. **6**, 685 (1953).
[56] Many of the low-angle x-ray diffraction studies were carried out in cooperation with Dr. A. Posner, Am. Dental Assn., Research Division, National Bureau of Standards, and with Dr. L. Bonar, Department of Biophysics, Massachusetts Institute of Technology.
[57] R. S. Bear, Advances in Protein Chem. **7**, 69 (1952).
[58] O. E. A. Bolduan, T. P. Salo, and R. S. Bear, J. Am. Leather Chem. Assn. **46**, 124 (1951).
[59] K. Kühn, W. Grassmann, and U. Hofmann, Naturwissenschaften **44**, 538 (1957).
[60] W. Grassmann, U. Hofmann, K. Kühn, H. Hörmann, H. Endres, and K. Wolf in *Connective Tissue*, Symposium organized by Council for International Organizations of Medical Sciences, R. E. Tunbridge, editor (Blackwell Scientific Publications, Oxford, 1957), p. 157.
[61] D. Turnbull, J. Appl. Phys. **21**, 1022 (1950).
[62] D. Turnbull, J. Chem. Phys. **20**, 411 (1952).
[63] D. Turnbull and B. Vonnegut, Ind. Eng. Chem. **44**, 1292 (1952).
[64] V. DiStefano, W. F. Neuman, and G. Rouser, Arch. Biochem. Biophys. **47**, 218 (1953).
[65] M. Polonovski and P. Cartier, Compt. rend. **232**, 119 (1951).
[66] R. J. Block, M. K. Horwitt, and D. Bolling, J. Dental Research **28**, 518 (1949).
[67] R. J. Block, Ann. N. Y. Acad. Sci. **53**, 608 (1951).
[68] D. B. Scott, Ann. N. Y. Acad. Sci. **60**, 575 (1955).
[69] H. R. Perkins and P. G. Walker, J. Bone and Joint Surg. **40B**, 333 (1958).
[70] J. Shatton and M. Schubert, J. Biol. Chem. **211**, 565 (1954).
[71] S. M. Partridge and H. F. Davis, Biochem. J. **68**, 298 (1958).
[72] K. Meyer, E. Davidson, A. Linker, and P. Hoffman, Biochim. et Biophys. Acta **21**, 506 (1956).
[73] M. A. Logan, J. Biol. Chem. **110**, 375 (1935).
[74] M. Schubert and D. Hamerman, J. Histochem. Cytochem. **4**, 159 (1956).
[75] J. Einbinder and M. Schubert, J. Biol. Chem. **185**, 725 (1950).
[76] H. P. Gregor, Trans. New York Acad. Sci. Ser. II, **18**, 667 (1956).
[77] E. S. Boyd and W. F. Neuman, J. Biol. Chem. **193**, 243 (1951).
[78] D. Burton and R. Reed, Discussions Faraday Soc. **16**, 195 (1954).
[79] D. Carlström, A. Engström, and J. B. Finean in *Fibrous Proteins and Their Biological Significance*, Symp. Soc. Exptl. Biol. **9**, 85 (1955).
[80] R. A. Robinson and D. A. Cameron, J. Biophys. Biochem. Cytol. **2**, Suppl. 253 (1956).
[81] S. M. Clark and J. Iball, Nature **179**, 94 (1957).
[82] S. M. Clark and J. Iball, Progr. in Biophys. and Biophys. Chem. **7**, 225 (1957).
[83] G. Wallgren, Acta Paediate. **113**, Suppl. (1957).
[84] T. W. Speckman and W. P. Norris, Science **126**, 753 (1957).
[85] H. Fernández-Morán and A. Engström, Biochim. et Biophys. Acta **23**, 260 (1957).
[86] M. L. Watson and R. A. Robinson, Am. J. Anat. **93**, 25 (1953).
[87] Author's unpublished results.
[88] C. Gregoire, J. Biophys. Biochem. Cytol. **3**, 797 (1957).
[89] T. Tsujii, D. G. Sharp, and K. M. Wilbur, J. Biophys. Biochem. Cytol. **4**, 275 (1958).
[90] N. Watabe, D. G. Sharp, and K. M. Wilbur, J. Biophys. Biochem. Cytol. **4**, 281 (1958).
[91] J. W. Gibbs, *The Collected Works of J. Willard Gibbs* (Longmans, Green and Company, New York, 1928), Vol. I.
[92] J. Frenkel, *The Kinetic Theory of Liquids* (Clarendon Press, Oxford, 1946, and Dover Publications, New York, 1955).
[93] M. Volmer and A. Weber, Z. physik. Chem. **119**, 277 (1926).
[94] M. Volmer, Z. Elektrochem. **35**, 555 (1929).
[95] M. Cohen, Transactions Metallurgical Soc. AIME **212**, 171 (1958).
[96] V. J. Schaefer, Chem. Revs. **44**, 291 (1949).
[97] B. Vonnegut, J. Appl. Phys. **18**, 593 (1947).
[98] M. Telkes, Ind. Eng. Chem. **44**, 1308 (1952).
[99] F. C. Frank, Advances in Phys. **1**, 91 (1952).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

43
Structure of Muscle Cells*

H. STANLEY BENNETT

Department of Anatomy, University of Washington School of Medicine, Seattle 5, Washington

IT IS my task to talk about the microscopic aspects of muscle structure. Muscle cells are specialized structurally for the function of contractility. In a muscle fiber, the ratio between deforming force and elongation can change rapidly and reversibly. This change characterizes the phenomenon of contractility. Contractility involves changes in the viscoelastic behavior of some fibrous protein structures in the muscle. In the relaxed state, a muscle fiber subjected to tensile stress along its axis will elongate, and a characteristic stress-strain diagram can be prepared. If a corresponding curve is taken while the muscle is contracted, the slope of the curve is much steeper. The shift from one elastic state to another can occur in a few milliseconds.

Contractile structures of this sort are found in a great many different forms. Inoué (p. 402) discusses certain features of contractile mechanisms other than those called muscle. Now, consider primarily a specialized class of biological contractile mechanisms called muscle. The boundary between muscle and nonmuscle is often not easy to draw. There are protozoa, for example, which have contractile filaments within certain portions of their complicated cells. Many of these contractile structures have properties of muscle.

Metazoans may likewise contain such dual-purpose cells. For example, some of the cells which surround the blood vessels of the earthworm serve the functions both of lining the channel and of controlling the size of the lumen. These endothelial cells contain contractile filaments which resemble those of muscle cells elsewhere in the organism. As another example, in hydra, a small metazoan organism, certain cells lining the gut cavity have an important role in the digestion and in the absorption of the food consumed. A part of these cells contain contractile filaments, and this part could be regarded as being similar to muscle. One might cite many other examples of dual or triple function cells, one of whose properties is that of contractility.

But even the well-defined muscle cells found in the limbs of a vertebrate organism may display more than one physiological function. In addition to the main function of contractility, the muscle fibers of the mammals have a very important role in the production of heat. A good fraction of the body temperature is generated by combustion in the muscle cells. If the heat losses from the body are sufficient, a very special type of contractile response known as shivering occurs. This is characterized by very rapidly repeated contractions which do not accomplish any useful work but which generate a good deal of heat, thus helping to restore a satisfactory thermal balance. In addition, there are very important electrical characteristics of muscle. The phenomenon of excitation of muscle is accompanied by electrical changes. These electrical changes involve voltage transients of 20 to 100 mv, depending on the muscle. These are generally only of local importance as far as that muscle is concerned. But, in an aqueous medium, these electrical transients can escape into the surrounding fluid and be detected by sensitive electrodes. In electric fishes, the muscle becomes modified so that the electrical transients become very highly directional and strengthened. In most cases, the contractility of such specialized muscle fibers is lost. In such a fish, each modified muscle cell is polarized, forming an electroplax. Considerable numbers of these electric generating units may be stacked up in series and arranged in parallel. In the most highly developed forms, the electric organ can produce formidable power. The torpedo ray, for example, can deliver a current of 60 amp at 80 v as a result of series and parallel summation of many of these specialized muscle cells. In the case of the electric eel (*Electrophorous electricus*), more of these cells are summed in series and potentials can achieve values of 300 or more volts. Currents in this case, with fewer generating units in parallel, are of the order of 0.3 to 0.5 amp. Thus, by suitable electrical coupling (by giving a suitable vector to the electrical discharge of the muscle and connecting the modified muscle units in series and in parallel), rather astonishing electrical effects can be produced. In addition, many fish have very sensitive electric sensing organs and can broadcast electric pulses (some of them small) to explore the surrounding water. By sensing these pulses with their special electric receivers, they test the conductivity of the water and use this information for navigation, for escaping enemies, and for finding prey.

Muscle fibers proper contain not only a contractile mechanism itself, but also other structures which control and fuel the contractile devices and render the whole apparatus a workable and effective part of the organism. The unit which accomplishes all of these coordinated physiological functions is a muscle fiber. These are really muscle cells, but since many of them are long and thread-like, they are called fibers. There are many kinds of muscle fibers. They vary from 10 to 100 μ in diameter and from 20 μ to several centimeters in length. The smaller muscle fibers usually contain only one nucleus.

* The preparation of the material presented in this report was assisted by Grant B-401 from the National Institutes of Health, Public Health Service, U. S. Department of Health, Education, and Welfare, and by a grant from the Muscular Dystrophy Associations of America, Inc.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

Many of the larger fibers, however, contain very many nuclei, up to hundreds.

Muscles also vary greatly with respect to the fraction of their resting length by which they can shorten. At one extreme are the flight muscles of insects, such as those of the fly or wasp. The change in length between the relaxed state and contracted state in this type of muscle is relatively small, of the order of 2 to 5%. At the other extreme, one might cite the proboscis retractor muscles in certain marine worms, where the changes upon shortening are of the order of 90% of the original length. In vertebrate skeletal muscle such as is found in humans, the changes in length between the contracted and relaxed state do not exceed 20% under ordinary circumstances.

For over a century, it has been realized that an explanation of contraction of muscle depended on an understanding of its molecular components and in a characterization of the actions and interactions of these components in terms of the contractile process. This concept of the participation of molecular units in contraction was explicitly stated in 1859 by Kühne,[1] when he made the first attempt to obtain from muscle proteins which might be important in the contractile process. He pressed out muscle juice and obtained from it a crude protein preparation which he called "myosin."

In recent years, with improvements in chemical and physical techniques, there have been a number of attempts to relate the molecular components of muscle with the ultrastructural anatomy and the actual process of contraction. Two important hypotheses have been proposed (see Morales, p. 426):

The first may be called the one-filament hypothesis, and the second the two-filament hypothesis. In the first case, contraction is considered to be the result of a change in the configuration of a unit consisting of a single filament. Shortening of many of these unit filaments in parallel and in series would occur when a muscle cell contracts. So far as I am aware, the clearest and earliest explicit statement of this hypothesis can be laid to Meyer.[2] He proposed a number of models in

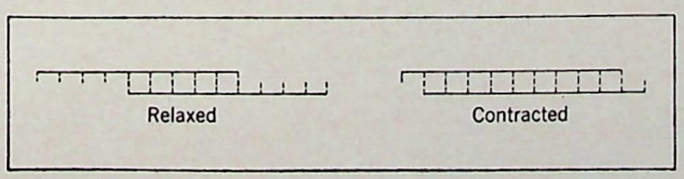


FIG. 1. Diagram of a conceptual model of the simplest possible muscle, based on the two-filament hypothesis of Huxley and Hanson. Each filament, represented by a horizontal solid line, is fitted with a series of binding sites, shown as short vertical dashed lines. The filaments constituting a pair overlap to a certain extent, and are bound to each other by a number of binding groups. In the relaxed state, the length of the overlapping portions is relatively short, and only a few of the binding sites are in contact. In contraction, the interaction between the binding sites changes, more binding sites become engaged, and one filament is drawn up alongside the other, increasing the proportion of overlap, shortening the total length of the system, and exerting tension in a direction parallel to that of the filaments. The filaments are of molecular dimensions.

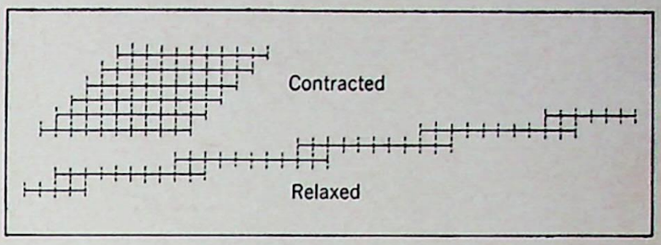


FIG. 2. Diagram of a conceptual model of a modification of the two-filament hypothesis designed to account for a muscle which can contract to a small fraction of its resting length. The concept holds that in the relaxed state a series of myofilaments, each fitted with two sets of bonding sites capable of interacting with those of fellow filaments, overlaps a neighboring filament to a slight extent at each end. Upon contraction, the length of overlap between adjacent filaments is caused to increase as each filament is pulled lengthwise along its neighbor by a change in the interaction between binding sites. The assembly thus behaves like a multiple set of extension ladders. The terminal extremities of the two terminating members of the series of extended relaxed filaments, however far separated, can be pulled by the contractile process to a distance apart limited only by the length of the longest filament of the series.

which attraction and repulsion of charged groups on a helix were responsible for contraction. Astbury,[3] from his x-ray diffraction work, elaborated on this concept. The view that the unit of contractility resides in a single filament dominated the literature until very recent years.

The second hypothesis involves the view that the unit structure in contractility consists of two parallel, overlapping and closely spaced filaments. One of the filaments moves parallel to the other during the contractile process. Thus, the over-all length of the systems is decreased, although each of the individual filaments maintains its original length (Fig. 1). This two-filament hypothesis, which has been most explicitly stated by Hanson and H. E. Huxley[4] and H. E. Huxley,[5] has been considered from the energetic and physiological point of view rather elaborately by A. F. Huxley.[6] It is now an important topic of discussion amongst muscle physiologists. This two-filament hypothesis can be regarded as a reversible special version of the concept of aggregation of macromolecules and protofibrils which F. O. Schmitt discusses in connection with collagen (p. 349). He points out that there is evidence that collagen macromolecules can aggregate in various spatial relationships with respect to one another. As applied to muscle, the theory would state that the protein filaments in muscle could shift reversibly from one position of parallel alignment to another.

According to this hypothesis, the simplest muscle would consist of two filaments (Fig. 1). An elaboration of this would permit one to build an analogous model to fit a fiber which contracted to only a small fraction of its relaxed length. As mentioned previously, such muscles are found in some invertebrates. Such a muscle might be composed of a series of filaments arranged in staggered array, as in Fig. 2. Upon contraction, these filaments would simply slide up on each other like the units of a multiple or extension ladder. This provides a way in which the two-filament hypothesis can be elabo-

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

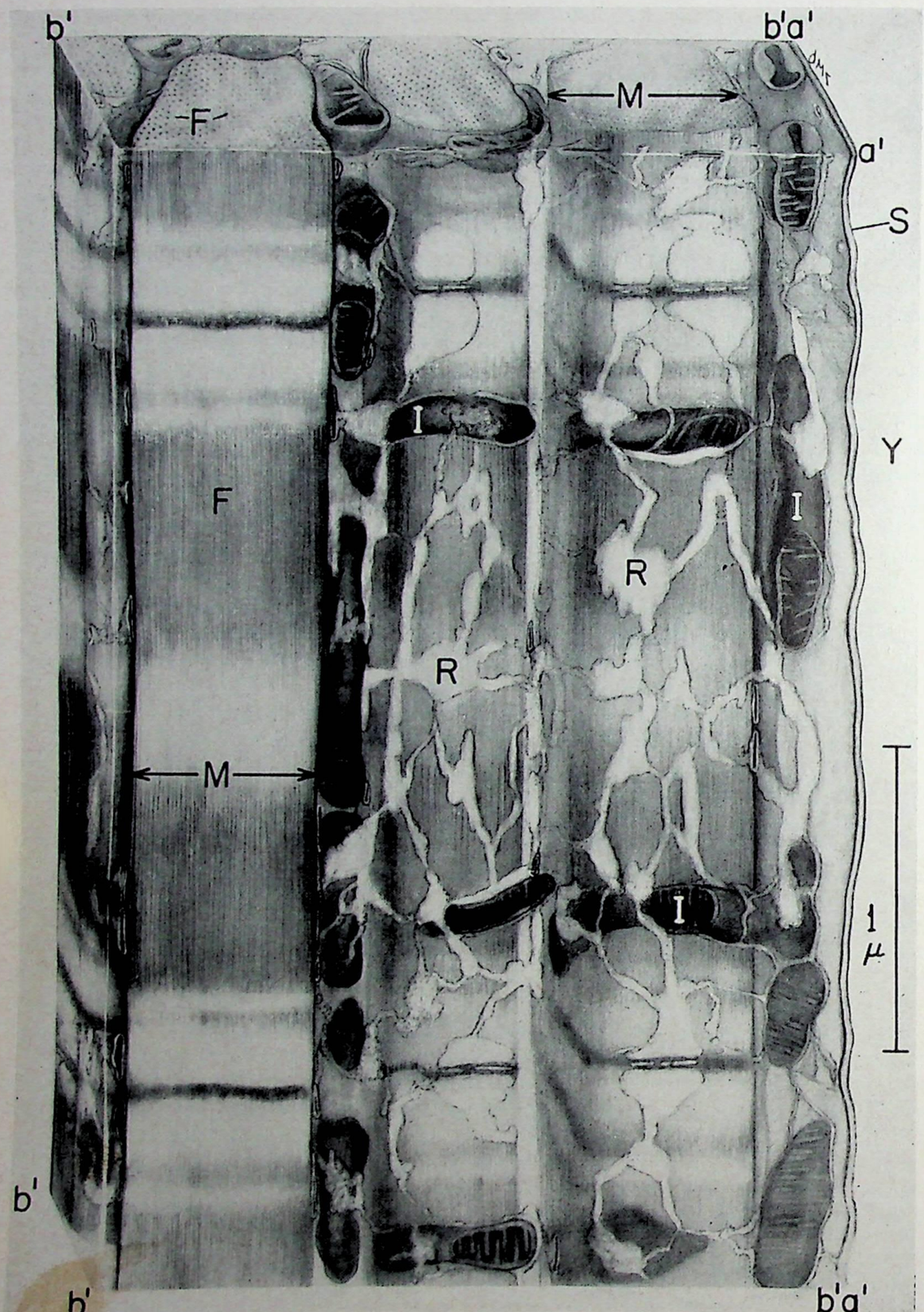


FIG. 3. A diagram to show some of the structural features of a typical striated muscle fiber, such as might be found in the limb muscle of a vertebrate. The figure represents a semitransparent block cut out of the muscle. *Y* designates extracellular space. The sarcolemma *S* is the plasma membrane of the muscle cell. Many mitochondria *I* are interspersed between the myofibrils *M*. Within the myofibrils are the myofilaments *F*, which are of molecular dimensions. Forming laceworks surrounding the myofibrils are elements of the sarcoplasmatic reticulum *R*, also called sarcotubules.

rated to account conceptually for some of the more extreme examples of muscular contraction. There is no experimental evidence, however, to support this model for muscles which contract to only a small fraction of their resting length.

On the basis of electron micrographs, the two-filament hypothesis is supported by the visualization in striated muscle of two sets of filaments.[5] In the muscles which Huxley studied, one type of filament is thicker than the other (Figs. 4 and 5). This is not an essential part of the concept, however, and it would not detract from the theory if, in some muscles, all of the filaments were found by electron microscopy to have geometrically similar dimensions.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

In addition to the contractile elements, muscles contain a number of other important components. Some of these are represented for one type of muscle in Fig. 3. The muscle is surrounded by a membrane which corresponds to the plasma membrane of the muscle and is called the sarcolemma (*S*). The concentrations of constituents of the fluid encompassed by the plasma membrane differ from those outside. If one places an electrode inside the muscle cell and connects it through a suitable recording device to an electrode in the fluid outside the muscle cell, one finds a potential difference between the inside and outside of the muscle. This potential difference has been related to the difference in ionic concentrations across the sarcolemma. Fairly satisfying relationships, based on the differential concentrations of potassium and sodium, have been derived to account for this potential difference. In electron micrographs, the sarcolemma (the structure responsible for separating these two ionic compartments) is of the order of 100 A thick. A measured 100-mv potential difference between the inside and the outside of the muscle fiber means that there is a voltage gradient of the order of 100 000 v/cm across the sarcolemma.

Katz (p. 466) presents some of the features of membranes of this sort in his discussion of nerves and neural conduction. If one applies a quick transient electric pulse to certain portions of the sarcolemma of many muscles, one gets an electrical disturbance which sweeps in all directions over the muscle fiber without loss of amplitude. Such electrical excitability resides in large portions of the sarcolemma, but is not found in the immediate region where nerves reach the muscle, nor is it found in the region where the tendons attach to the muscle fiber. From this difference in electrical behavior, one can conclude that the molecular organization of the sarcolemma varies in different portions. There is, however, no satisfactory correlation of these presumed differences in molecular structure with images of the sarcolemma in electron micrographs, nor is it known in any detail what these differences might be physiochemically.

The sarcolemma can be traced continuously around both ends of the muscle. In general, in most vertebrate muscle and in some insect muscle, the fibers of the tendons are bonded in some way to the outer surface of the sarcolemma. Certain noncontractile intracellular filaments continuous with the myofilaments also connect with the sarcolemma, but with its inner surface, opposite the tendon attachments. This arrangement is capable of transmitting the tension generated within the fiber to some extrafibrillar unit, such as bone.

Most muscle fibers, but not all, are equipped with a nerve supply which, in more elaborate fibers, serves a number of different functions. In some types of vertebrate muscle, such as skeletal muscle, there is a main motor nerve which supplies each fiber. When this nerve is excited, the muscle responds by contracting. In addition, in many cases there are sensory organs located in the tendons. These sense organs are similar to strain gauges in that they can respond to changes in tension and send impulses to the central nervous system. These signals can then be processed and correcting directions relayed back to the muscle fiber. This provides a type of feedback loop which regulates the strength and temporal nature of the muscular contraction. In certain muscles, much more elaborate types of regulatory devices are present. One is known as a muscle spindle. This is a complex structure embedded amongst ordinary muscle fibers and consisting of a very delicate muscle fiber inside a connective tissue capsule. This special fiber is fitted with a nerve supply containing motor and sensory components. The motor component can actuate the contractile mechanism of the very fine fiber within the spindle capsule and thus vary the bias on this sensory device. The sensory fibers reach nerve endings associated with the fiber and pick up signals which result from changes in tension. These are sent to the central nervous system. These examples do not exhaust the types of regulatory mechanisms which operate in muscle fibers. In certain cases, rather different mechanisms are found.

One can divide the contents of the muscle cell into two components: the contractile portion and the noncontractile portion. The latter is called the sarcoplasm. This sarcoplasm contains mitochondria, nuclei, fat droplets, glycogen, and other structures. The fat and the glycogen serve primarily as fuel. Among the other structures, muscle contains an elaborate internal system of membranes (Fig. 3, *R*). These have been studied by Sjöstrand,[7] by Andersson,[8] by Porter and Palade[9], and others. These membranes often define tubular structures. This has led Sjöstrand to speak of them as "sarcotubules." The structures were seen with the light microscope by Retzius (1881, 1890) many years ago and termed the "sarcoplasmatic reticulum."[10,11] Thus, one can choose between two terms, each proposed by an eminent Swedish anatomist.

This internal system of membranes has, in general, not enjoyed much attention. It has been ignored completely by biochemists, and only recently has received the gaze of physiologists. Indications at the moment are that it may have a role in transmitting an excitatory signal from the sarcolemma to the contractile elements within the fiber.

Many of the features introduced above are diagramed in Fig. 3, which represents a semitransparent block cut out of a piece of striated muscle. It shows, in a rather over-simplified way, many of the features of muscle structure discussed in the foregoing. *Y* represents the intercellular space outside the fiber. The double lines *S* represent the sarcolemma or plasma membrane of the muscle. This has been severed by a knife in the directions a'–a'. The flesh itself has been cut along b'–b'. The myofilaments *F* are represented by the very small dots in the cross sections at the ends of the block and by the very fine striations seen on longitudinal section. In this

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

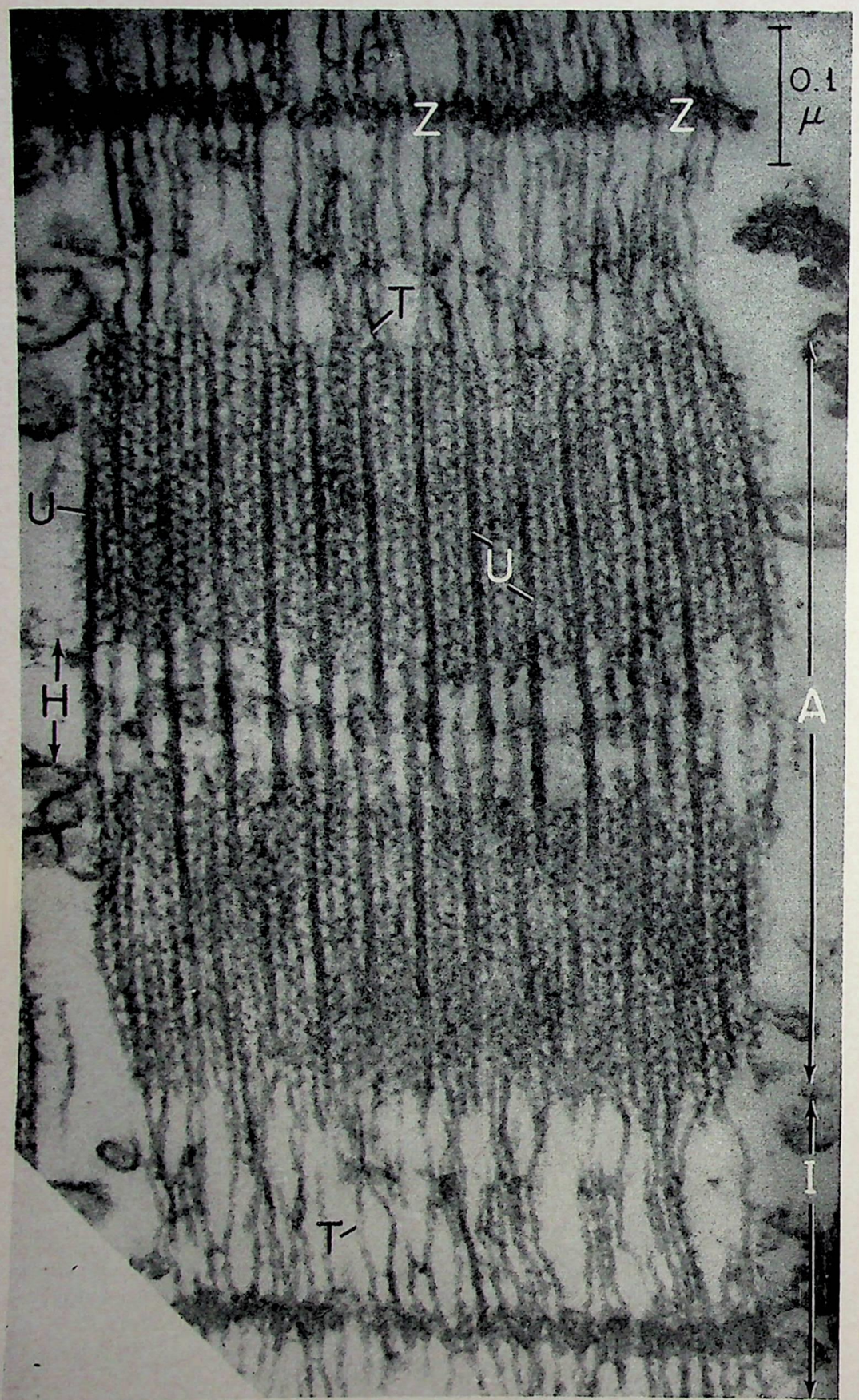


FIG. 4. An electron micrograph showing a single sarcomere from the striated muscle of a rabbit. This very thin section is cut nearly parallel to the axis of the myofilaments. Thin filaments *T* and thick filaments *U* can be distinguished and their overlap is displayed. *A*, *I*, *H*, and *Z* designate the corresponding cross bands of the striped muscle repeating unit, the sarcomere. Electron micrograph courtesy of H. E. Huxley.

muscle, the myofilaments are organized in groups, the myofibrils *M*. Myofibrils are found in many muscles but not in all. They are approximately 0.5 to 2 μ in diameter. In striated muscles, one finds a longitudinal repeating pattern, which F. O. Schmitt (p. 349) discusses earlier. This repeating pattern usually varies in relaxed muscle from about $1\frac{1}{2}$ to 2 μ. However, in invertebrate striated muscle, the length of the unit may extend to

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

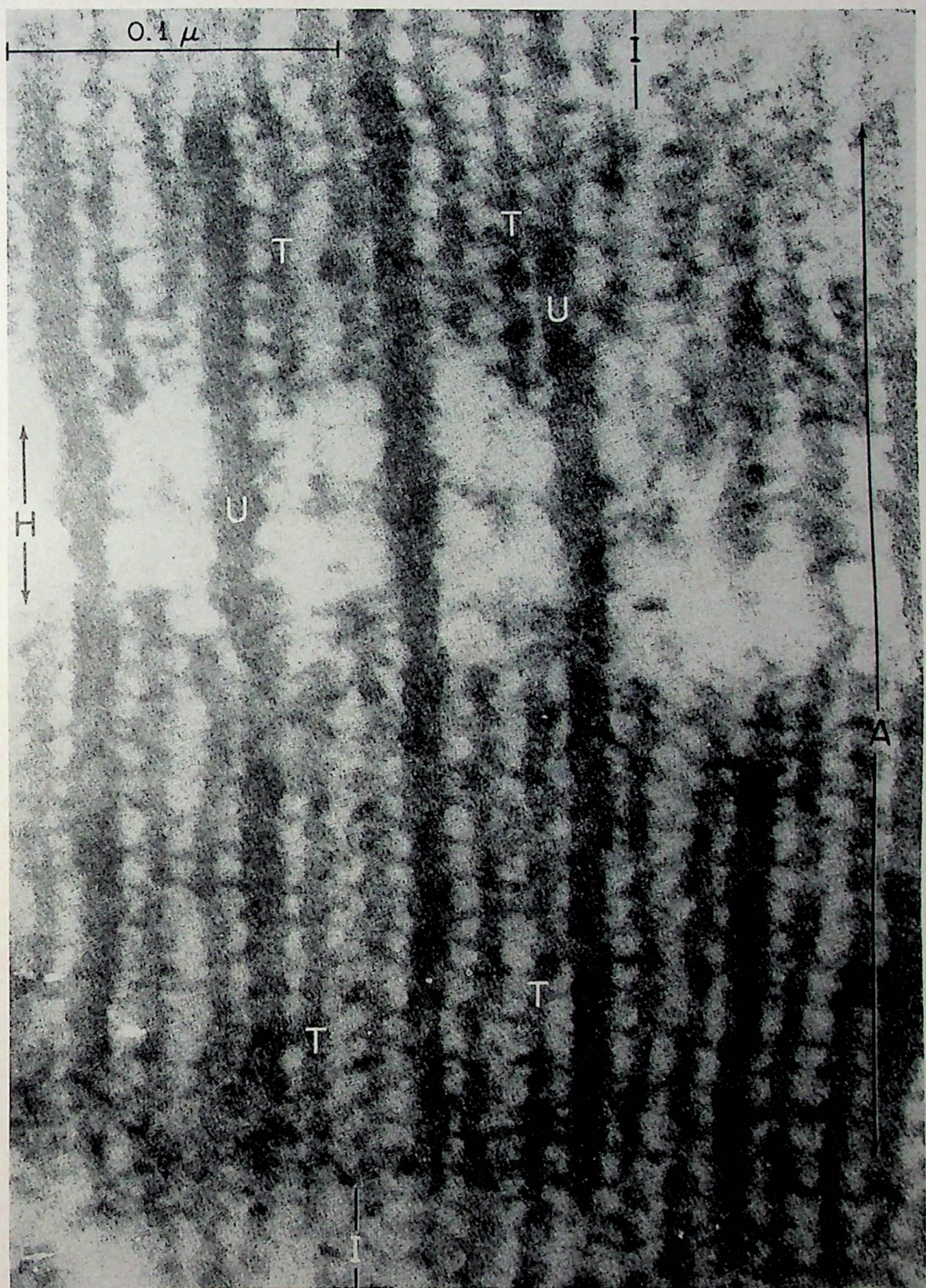


FIG. 5. A high-power detail of a portion of Fig. 4. The thin filaments *T* can be seen interdigitating with the thick filaments *U*, overlapping throughout the *A* band, except for the region of the light *H* band in the center. Slender bridges can be seen connecting the thick and thin filaments. These are believed to represent interaction sites, such as those represented by vertical short dashed lines in Figs. 1 and 2. Electron micrograph courtesy of H. E. Huxley.

10 or 15 μ, and in certain extreme cases (in certain unusual worms) repeating patterns of 50 or more μ can be encountered.

In most muscle, one finds many glycogen granules scattered about in the sarcoplasm. These are not shown here. In addition, particularly in certain phases of

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

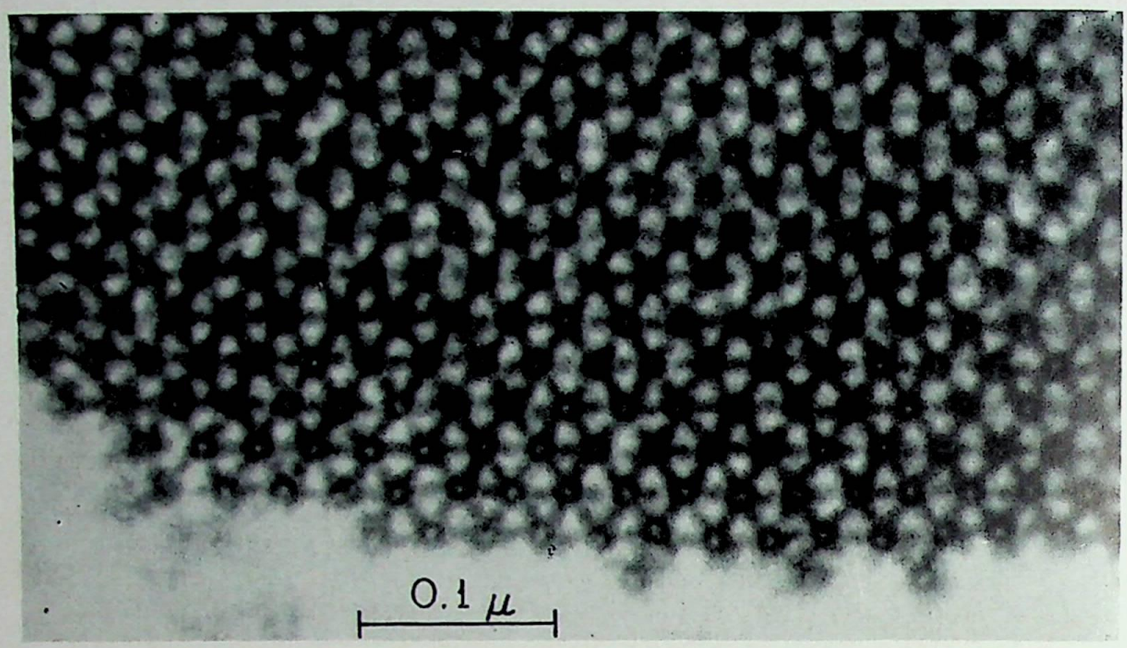


FIG. 6. An electron micrograph showing a cross section of a portion of a myofibril of the muscle of a fly. The thick myofilaments appear in cross section as circles arranged in hexagonal array, connected to each other by bridges forming equilateral triangles. In the middle of each bridge is a dense spot which may represent cross sections through thin filaments. Electron micrograph courtesy of A. J. Hodge.

physiological activity, one can find some intramuscular fat. If the animal is starved somewhat, a good deal of fat will appear in heart and other muscles. Under those conditions, mitochondria are closely associated with the fat droplets as though they were burning the lipid as fuel. In migrating salmon and in migrating geese, which feed extensively before their long journeys, the muscle fibers and adjacent cells accumulate very large amounts of fat. It is thought that the distance which a flock of geese can cover in a single flight is determined by the amount of fuel they can carry. It is much more economical to carry it in the form of fat than in the form of glycogen, since they can burn the hydrogen in the fat.

Turning back to Fig. 3, the mitochondria *I* can be clearly seen. Close to them and intertwining with the myofibrils are elements of the sarcoplasmatic reticulum of Retzius, *R*. In certain places, it can be seen that this reticulum comes in close contact with the mitochondria. It forms an extensive plexus here and there within the sarcoplasm of the muscle fiber. Thus, in gross outline it forms a network appearance. This gave rise to the concept that it was a reticulum. But if one cuts it in cross section, one sees that it is composed of membranes which form tubules which are very often flattened. Sjöstrand's term of "sarcotubule" is also appropriate, as it describes a second important feature of the structure. Retzius suggested in 1881 that this reticulum might pick up a signal from the sarcolemma and convey that signal deep into the interior of the fiber, thus exciting myofibrils a long way from the surface. Experimental justification of this concept, however, was a long time in coming. Only within the last two or three years A. F. Huxley and Taylor (see A. F. Huxley[6]) have brought forth evidence that there is indeed a conducting pathway which conforms to these specifications. Huxley and Hodgkin are now exploring the possibilities that the conduction might in fact reside within this system of tubules or membranes.

Robertson[12,13] has studied the sarcolemma in the electron microscope at high resolution. He has observed that it conforms in structure to the "unit membranes" found in other cells, presenting two peaks in density about 55 A apart, separated by a region of lesser density. On the outer surface of the sarcolemmatic unit membrane is a cloud of material which gives the chemical reactions associated with polysaccharides. It is believed to contain some glycoproteins. This type of polysaccharide-rich coating is found very frequently in association with plasma membranes, so its presence in association with the sarcolemma of muscle is not exceptional.

There is reason to think that there are probably two layers of lipid and some protein in the unit membrane comprising the sarcolemma. Various attempts have been made to fit these layers into the patterns of density seen in Robertson's micrographs. As Sjöstrand and Hodge point out, precise information about this point is not available. My own view is that all of the models concerned with the actual disposition of lipid within these membranes are so uncertain that one of our major tasks is to make a detailed map of the true arrangements of the proteins and lipid in biological membranes. I am doubtful of the usual presentation of the lipoprotein membrane structure as presented by Davson and Danielli,[14] where the lipid molecules are presented with their polar groups out and their nonpolar groups in.

Figure 4 is one of H. E. Huxley's elegant electron micrographs representing the arrangements of myofilaments in a myofibril. It provides some of the evidence for the hypothesis there are two interacting types of filaments differing from each other in structure. One, the thin filament *T*, is envisioned as starting at the *Z* band and extending a certain distance. The second type, the thick filament *U*, occurs in the middle of the

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

repeating unit, which is called the sarcomere. Contraction is envisioned as involving an interaction between these two filaments, so that traction is exerted and the ends of the large filaments approach the *Z* band. Figure 5 shows the interdigitation of these filaments at higher power. In cross sections, the large filaments are arranged in a hexagonal pattern, and the small filaments are arranged in special relationship to them. In muscle from the rabbit, the small filaments are arranged so that each one is located in the center of a triangle formed by the large ones. One can see that, in such a case, if a section is cut through the muscle sufficiently thin to encompass only one layer, a plane can be found which will accommodate the thick filaments and two thin ones. Figures 4 and 5 are examples of such sections visualized in the electron microscope. Of importance are certain bridges which can be seen transversely extending between the two types of filaments. These may represent sites of interaction between the two filaments.

Hanson and Huxley[4] have proposed that the thick filaments contain the protein myosin and the thin filaments contain the protein actin. Huxley interprets contraction of the muscle as a change in the bonding occurring through reaction sites at the bridges, taking place in such a way that the actin and myosin filaments slide along each other.

Figure 6 is an electron micrograph by Hodge.[15] It represents a cross section of insect muscle. The large filaments and the cross bridges connecting them in a hexagonal pattern can be seen clearly. There are no small filaments lying in the center of the triangles formed by the large filaments in this particular type of muscle. However, in many cases, one can see a dense spot in the middle of the line connecting adjacent large filaments. It can be shown that these represent cross sections of small filaments which, in this type of muscle, are arranged in the centers of lines connecting individual large filaments rather than in the middle of triangles bounded at the corners by the large filaments. Huxley, however, believes that this is a trivial difference and that in both cases the interactions between the small and large filaments are the basis of muscular contraction.

Whether or not one thinks of contractility as residing in one filament or in the interaction between two filaments, these micrographs of Huxley and of Hodge have gone far in showing the geometrical framework within which the contractile mechanism resides.

BIBLIOGRAPHY

[1] W. Kühne, Arch. Anat. Physiol. u. wiss. Med., Leipzig 564–642, 748–835 (1859).

[2] K. H. Meyer, Biochem. Z. **214**, 253 (1929).

[3] W. T. Astbury, Proc. Roy. Soc. (London) **B134**, 303 (1947).

[4] J. Hanson and H. E. Huxley, Symposia Soc. Exptl. Biol. **9**, 228 (1955).

[5] H. E. Huxley, J. Biophys Biochem. Cytol. **3**, 631 (1957).

[6] A. F. Huxley, Progr. in Biophys. and Biophys. Chem. **7**, 255 (1957).

[7] F. S. Sjöstrand, Intern. Rev. Cytol. **5**, 455 (1956).

[8] E. Andersson in *Electron Microscopy*, Proceedings of the Stockholm Conference, September, 1956 (Academic Press, Inc., New York, 1956), p. 208.

[9] K. R. Porter and G. F. Palade, J. Biophys. Biochem. Cytol. **3**, 269 (1957).

[10] G. Retzius, Biol. Untersuchungen **1**, 1 (1881).

[11] G. Retzius, Biol. Untersuchungen, Neue Folge, **1**, 51 (1890).

[12] J. D. Robertson, J. Physiol. **140**, 58 (1957).

[13] J. D. Robertson, Anat. Record **130**, 440 (1958).

[14] H. Davson and J. F. Danielli, *The Permeability of Natural Membranes* (Cambridge University Press, New York, 1943), p. 64.

[15] A. J. Hodge, J. Biophys. Biochem. Cytol. **1**, 361 (1955).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

44
Motility of Cilia and the Mechanism of Mitosis*

SHINYA INOUÉ

Department of Biology, The University of Rochester, Rochester 20, New York

THE movement of cilia and chromosomes are interpreted in terms of the microscopic and fine structures in cells. Section 1 discusses structure and motility of cilia; Sec. 2, microscopic structure of the mitotic spindle; Sec. 3, centers of organization in spindle and cilia; and Sec. 4, on the physicochemical nature of the spindle; and Sec. 5, relation of cilia and chromosome movement to muscle contraction.

1. STRUCTURE AND MOTILITY OF CILIA

Many small organisms—bacteria and protozoa, sperm and embryos of larger organisms—are propelled by beating their thin whip-like cilia or flagella (used interchangeably here). Also, ctenophores even a foot long and several inches wide can swim about or adjust their gravitational orientation by coordinated beating of their ciliary bundles. When the cell or organism is fixed, as in the gills of mussels and clams and in the human trachea, cilia create a current capable of pumping a considerable quantity of liquid or mucous material.[1–3]

The length of cilia may range from a few microns to several millimeters, and there may be one to several thousand cilia per cell but their diameter is quite constant, usually between a tenth and a half micron. Where a number of cilia occur in a row, adjacent cilia beat slightly out of phase with each other and a regular propagating wave is observed.

At the base of each cilium is found a characteristic bulbous enlargement, the basal granule. Further, proximal to the basal granule, rootlets sometimes are found which may function as anchorage or serve to conduct impulses. The membrane of the cilium appears continuous with that of the cell body.

Examined with an electron microscope (Fawcett and Porter,[4] and also dicussed in the following), cilia reveal a characteristic inner fibrillar pattern of amazing uniformity. Near the periphery of the cilium, there are usually nine (or occasionally a multiple thereof) fibrils, each composed of two filaments (or tubules?) some 100 to 200 A in diameter. Surrounded by the outer nine are two additional central fibrils of somewhat smaller diameters (Figs. 1 and 2). The nine outer fibrils reach the base of the cilium and appear to merge laterally into a hollow tube to make up the basal granule. The two central filaments also extend the whole free length of the cilium but apparently do not reach the basal granule. At the tip of the cilium, the outer nine and inner two fibrils are said to merge. In cilia with distinct directional beating, the line intersecting the two central filaments lies at right angles to the direction of beat.

The pattern of nine plus two has been found in cilia from a wide variety of cells, in fact, in practically all cilia observed with adequate resolution. The eleven fibrils are unlikely to be artifacts formed during preparation for electron microscopy, as sperm-tail flagella macerated in distilled water also show the frayed eleven fibrils in dark field illumination.

Given this structure, how does one explain the mechanism of ciliary beat? The pattern of beat may be relatively simple, as shown in Fig. 3(a). There is a recovery stroke in which the limp cilium stiffens from the base up, and an effective stroke where bending is mostly at the base and the rest of the cilium acts as if it were stiff. The same flagellate organism which swims forward by this

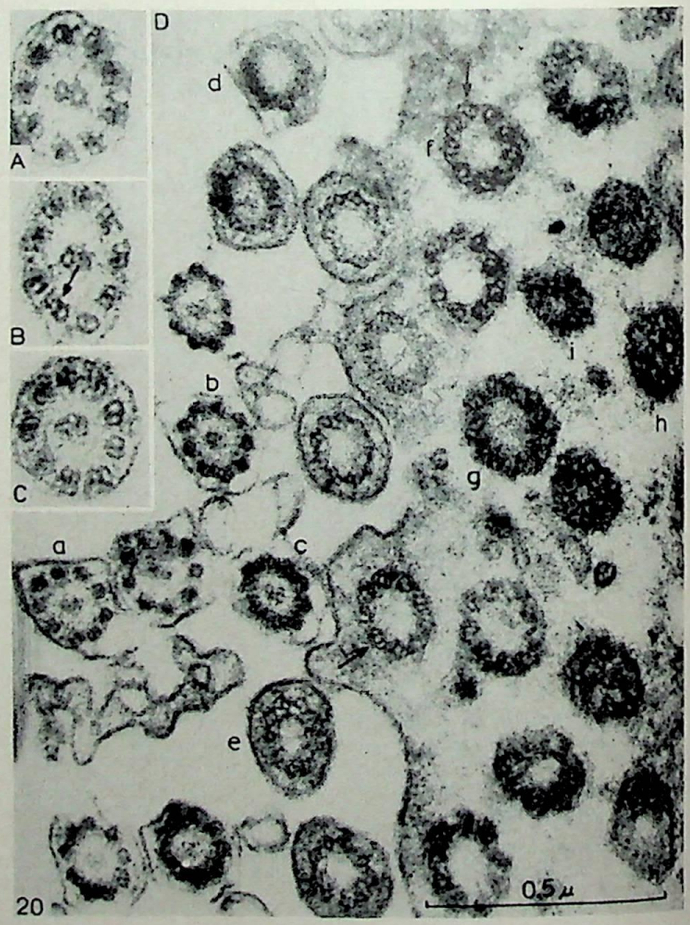


FIG. 1. Electron micrograph of cross sections of rat-trachea cilia. Compare with interpretive diagram, Fig. 2 [from J. Rhodin and T. Dalhamn, Z. Zellforsch. u. mikroskop. Anat. 44, 345 (1956)].

* Original work appearing in this paper was supported in part by the Public Health Service, U. S. Department of Health, Education, and Welfare (G-3002-C), and by the American Cancer Society.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

beat may swim backward, also sidewise or circularly, as shown in Fig. 3(b).[2,5] Bradfield[6] postulates that waves of contraction proceed along the length of the outer nine fibrils with a message perhaps traveling in advance along the two inner ones. The waves of contraction may be started at the base by a commutator-like device which would result not in a synchronous contraction but in waves with various phase lags. It is more likely, however, that the outer fibrils may be the conductive elements, the inner two at least partaking a more active function in beating. For, in certain sensory cells (see following) and at the embedded base of each cilium[4,7] where one would expect conduction but not contraction, one in fact finds the outer nine fibrils and not the inner two.

It was tacitly assumed in the foregoing that contraction of the fibrils was the basis for cilia beat. Actually, the only evidence for contraction of cilia components (at the molecular level) appears to lie in the x-ray diffraction studies of Astbury *et al.*[8] There they find in flagella collected from bacteria, in addition to an α-protein pattern (which is characteristic of many fibrous proteins, such as keratin, myosin, elastin, etc., and is believed to reflect the fundamental spacing of the polypeptide backbone), a folded β pattern. This supercontracted pattern they believe reflects the folding of a fraction of the polypeptide chains responsible for contraction. The protein isolated from this bacterial flagella preparation lacks sulfur-containing amino acids and appears dissimilar to any of the muscle components so far known.

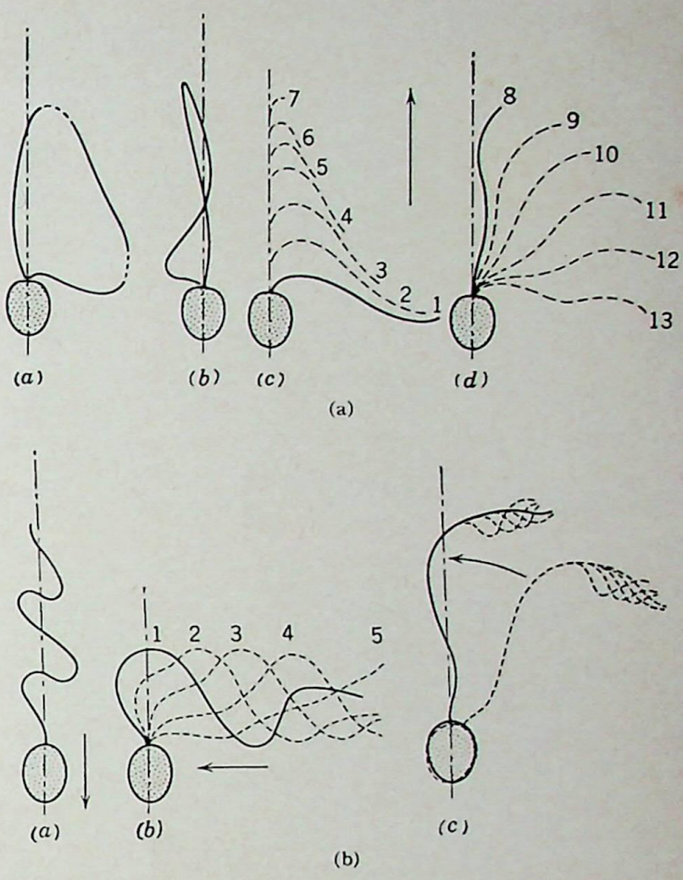


FIG. 3. Various patterns of beat of a *Monas* flagellum. Arrows indicate directions of swimming [from M. Hartmann, *Allgemeine Biologie* (Gustav Fischer Verlag, Stuttgart, 1953), fourth edition; after Krijgsman[5]].

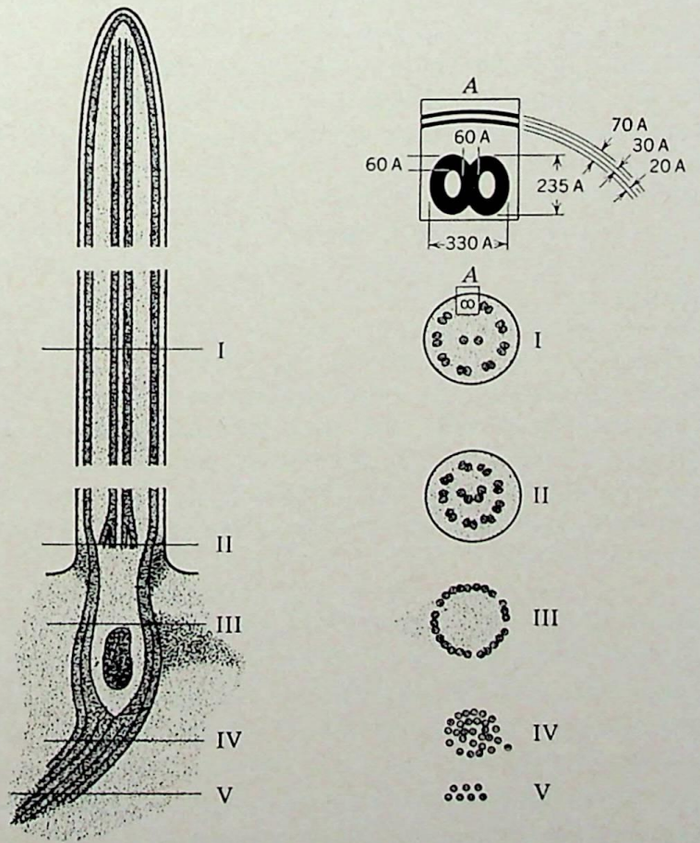


FIG. 2. Interpretation of electron micrograph (Fig. 1) showing fine structure of cilia [from J. Rhodin and T. Dalhamn, Z. Zellforsch. u. mikroskop. Anat. 44, 345 (1956)].

Aside from localized contraction, the mechanism of ciliary beat has been interpreted by other schemes also, such as local swelling, reciprocal pumping of a liquid in and out of the cilium, or discontinuous flow of material through the cilium.[1,3,9] The exact site of the motor function in the cilium is also disputed, but nevertheless there exist certain observations (Secs. 3 and 5) which link the structure and function of these minute structures to other specialized motile structures of the cell.

DeRobertis, Sjöstrand, and others[10,11] made an interesting discovery related to the structure of cilia in the filaments of the retinal-rod cells and of the sensory-hair cells of the inner ear. These fibrous elements, believed for some time to be derived from embryonic cilia, also show a fine structure similar to that of cilia described in the foregoing. In these apparently nonmotile fibers, the same outer nine fibrils are found, but the inner two are missing.

2. MICROSCOPIC STRUCTURE OF THE MITOTIC SPINDLE

Cilia may beat as frequently as a hundred cycles per second. Chromosomes, on the other hand, move extremely slowly, the maximum velocity being a few

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

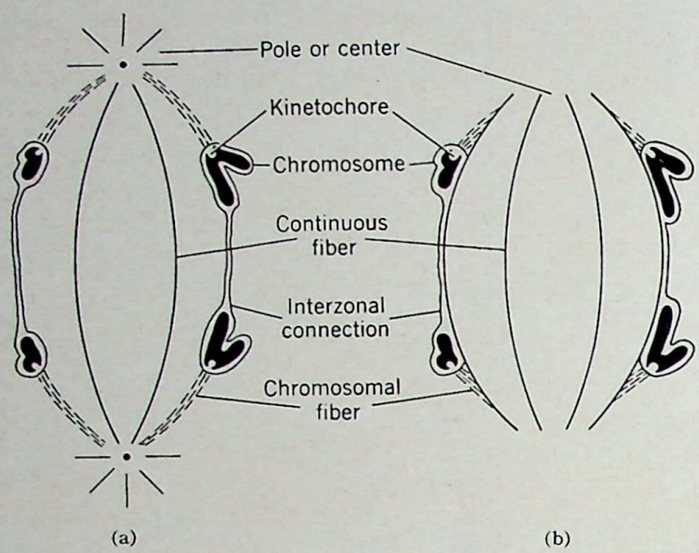


FIG. 4. Schematic diagram of mitotic spindle, (a) with centrioles and (b) without (modified from Schrader[19]).

microns per minute at anaphase (for reviews of mitosis see references 12 to 21). However, certain elements of the mitotic apparatus responsible for chromosome movement may be identical to, or at least have properties common to, portions of cilia. Before discussing this problem in Sec. 3, the structure of representative mitotic apparatuses are described.

Using Schrader's description,[19] the following terms are used. A dot or rod-like structure, the "centriole," is found at the poles of the mitotic spindle of many animal cells and occasionally in plant cells. The centrioles are morphologically the focal points for the spindle fibers and the astral rays [Fig. 4(a)]. In an average plant cell [Fig. 4(b)], both the centrioles and the astral rays are missing and the spindle fibers show less tendency to converge at the poles. Between the two spindle poles lie "continuous fibers." From the "kinetochores," a specific region of the chromosomes "chromosomal fibers" extend to or toward the spindle poles.

For half a century, the reality of these fibers in *living* cells has been disputed, for, with very few exceptions, they could be seen only in cells after fixation and staining. Recently, however, with improvements of the polarizing microscope, the author has been able to show these fibers clearly in living cells of many animals and plants by virtue of their positive birefringence (strength of birefringence $10^{-3}-10^{-4}$).[22,23] In animal cells, the fibers converge and their birefringence is stronger adjacent to the chromosome kinetochores and near the centrioles [Fig. 5(a)]. In plant cells, the situation is similar at the kinetochore region, but toward the "poles" the birefringence is weaker and the fibers are more diffuse [Fig. 5(b)]. During anaphase movement, the birefringence of the chromosome fibers persists and always is strongest adjacent to the kinetochores and the centrioles. The birefringence of the continuous fibers falls once and then rises again after complete separation of the chromosomes. The midregion containing fibers with strong secondary birefringence becomes the phragmoplast of plant cells [Fig. 5(c)]. Within the phragmoplast, small granules align, fuse, and become the cell plate, which divides the original cell into two. In animal cells, the cytoplasm generally cleaves inward at right angles to the spindle remnant forming two new cells with one nucleus each. Time-lapse motion pictures of these processes have been made by the author using a special polarizing microscope and were shown at the meeting.

The material of the mitotic apparatus has been isolated from sea urchin and other eggs by Mazia, Dan, and their collaborators in quantities sufficient for chemical analyses.[24–26] Amino-acid composition of their protein fraction (molecular weight *ca* 45 000) shows, unlike the bacterial flagella protein, a fair content of sulfhydryl-containing amino acids.

Although the spindle isolated after alcohol treatment by Mazia and Dan is stable, the fibers of the mitotic spindle in living cells are apparently extremely labile. The spindle fibers may disappear by slight mechanical agitation or by low-temperature treatment of the cell, only to reform in the course of a few minutes (see Sec. 5, also Carlson[13,27], Chambers,[28] Inoué,[29] and Östergren[30]).

3. CENTERS OF ORGANIZATION IN SPINDLE AND CILIA

Although the apparent velocity and very probably the stability of the spindle fibers differ by orders of magnitude from the cilia, the fibrous elements of the two structures may be formed or organized in a very similar fashion. The argument follows.

(A) The birefringence of the spindle fibers and astral rays is strongest adjacent to the kinetochores and centrioles throughout metaphase and anaphase (Inoué[23] and Schmidt[31]). The fibers are arranged radially (within restricted cones, in the case of the kinetochores) from these centers, and the growth of the spindle (at least in animal cells) takes place by lengthening of the fibers joining the centers. This strongly suggests that both the centrioles and the kinetochores are centers of fiber orientation.

(B) Growth of the axial filament of the sperm-tail flagellum starts from the basal granule, which same structure during the last mitosis acted as a centriole of the spindle.[21,32] In certain protozoa, flagella and the mitotic spindle both grow simultaneously from common giant centrioles.[14,32]

(C) In abnormal divisions of snail spermatocytes, some chromosomes lose their kinetochores and cannot partake in mitosis. The Pollisters[33] have shown that a clear correlation exists between the number of such chromosomes and the number of supernumerary basal granules, which migrate to the cell periphery and form the same number of extra sperm tails. Also, those kinetochores earlier dissociated from chromosomes form small extra astral rays while the spindle for the next division is formed.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

(D) With the electron microscope, centrioles of the mitotic apparatus have been shown to exhibit the same general structure and dimensions as that earlier described for the cilia basal granules, namely, a cylindrical structure containing nine groups of rods or tube-like elements.[34–37] This same structure is observed for the basal granule of sperm-tail flagellum.[38,39]

Thus, it appears that basal granules of cilia, kinetochores, chromosomes, and centrioles of the mitotic apparatus all act as centers of fibrous organization in cells. The basal granules, centrioles, and kinetochores, may be in fact identical structures, taking on different functions at different loci within the cell (see also Meves[40]).

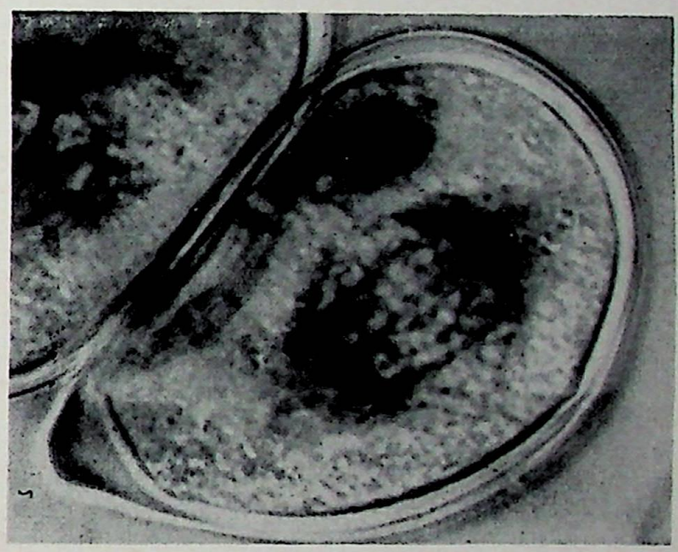
4. ON THE PHYSICOCHEMICAL NATURE OF THE MITOTIC SPINDLE

Electron-microscope studies have revealed a characteristic fine structure in cilia (Sec. 1). The possible identity of their basal granules to centrioles of the spindle also was strengthened (Sec. 3). However, this technique as yet has revealed rather little of the structure and behavior of spindle fibers and kinetochores.[25,41,42] The lack of success, I believe, is attributed to the difficulty or impossibility of preserving the spindle material in a reasonably native form after fixation and electron bombardment. In contrast, the polarizing microscope enables one to observe birefringence of spindle fibers in actively dividing cells (Secs. 2 and 3). Although the fine structure cannot be resolved directly, measurement of birefringence allows one to interpret the changes taking place in the microscopically unresolvable domain. This section describes further polarization-optical observations which may shed some light on the physical chemistry of the mitotic spindle.

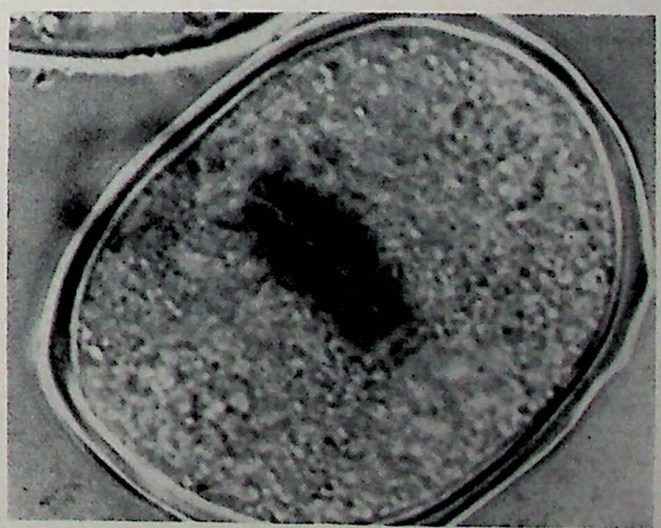
The spindle of the egg cell (of a marine worm *Chaetopterus*) can be stretched if the cell is flattened very gently. The length of the spindle is then found to be strictly proportional to the diameter of the compressed egg.[22] This relation is explained by the attachment of the spindle poles through astral rays[43,44] to the cortical-gel layer. Immediately upon stretching, the spindle is thinner and more pointed at the poles, while in a minute or two it grows fatter and the birefringence increases. When the egg is compressed suddenly, the link between spindle poles and the cortical gel is apparently broken and the spindle shortens as it loses its birefringence (Fig. 6).

At a given length, the birefringence of the spindle fibers is a function of temperature.[29] With abnormally low temperature (4° to 6°C), the spindle birefringence is abolished completely. When the temperature is raised, the birefringence returns. The loss of birefringence with low temperature is rapid (less than half a minute), but when the temperature is raised the birefringence and structure of the spindle fluctuate until, after several minutes, they reach an equilibrium specific for the new

(a)



(b)



(c)

FIG. 5. Birefringent spindle fibers in living cells. Photographs are printed as negatives and show spindle fibers parallel to spindle axis black. (a) *Chaetopterus pergamentaceous* metaphase; (b) *Lilium longiflorum* early anaphase; (c) the same, phragmoplast with early cell-plate formation (modified from Inoué[23]).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

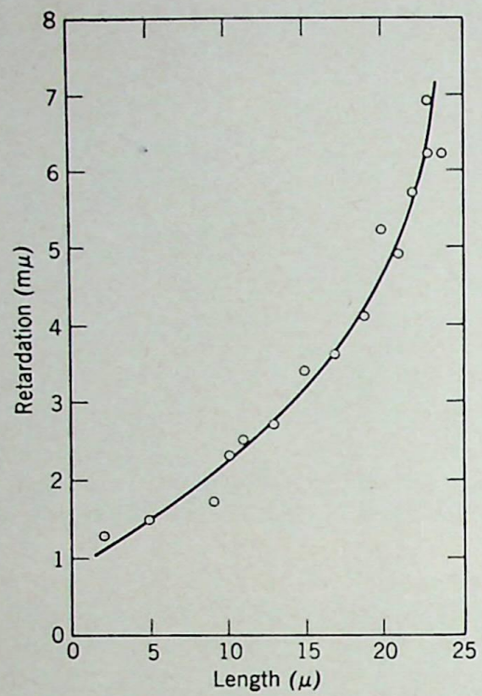


FIG. 6. Retardation of *Chaetopterus* spindle plotted against length [from S. Inoué, J. Exptl. Cell Research Suppl. 2, 305 (1952)].

temperature. The native spindle is, therefore, in a temperature-sensitive equilibrium.

The equilibrium birefringence at various temperatures is plotted in Fig. 7. If it is assumed that the birefringence of spindle fibers is directly proportional to the amount (B) of material oriented in that region, that only the equilibrium constant [$k(T)$] between oriented and nonoriented material is influenced by temperature, and that the total amount (A_0) of the orientable material in the same region remains constant, then the equilibrium is expressed by

$$A_0 - B \overset{k(T)}{\rightleftarrows} B.$$

A_0 was assumed constant since the spindle in the *Chaetopterus* egg is in metaphase equilibrium and also because cells which already have entered metaphase can go through division even in the presence of metabolic inhibitors such as cyanide and carbon monoxide.[45]

A_0 then is determined as the asymptote of the curve in Fig. 7. Were these assumptions warranted, one should expect a linear relationship between log $B/(A_0-B)$ and $1/T°K$. Figure 8 shows this plot. From the slope and intercept we (Morales[46] and Inoué) calculate the evolution of 28 kcal of heat per mole reacted, while at 25°C a free-energy change of -1.8 kcal/mole and an entropy increase of 100 eu/mole is observed.

The very low free-energy change agrees with the proposed lability of the spindle structure (weak gel with small number of active hydrogen bonds?), while

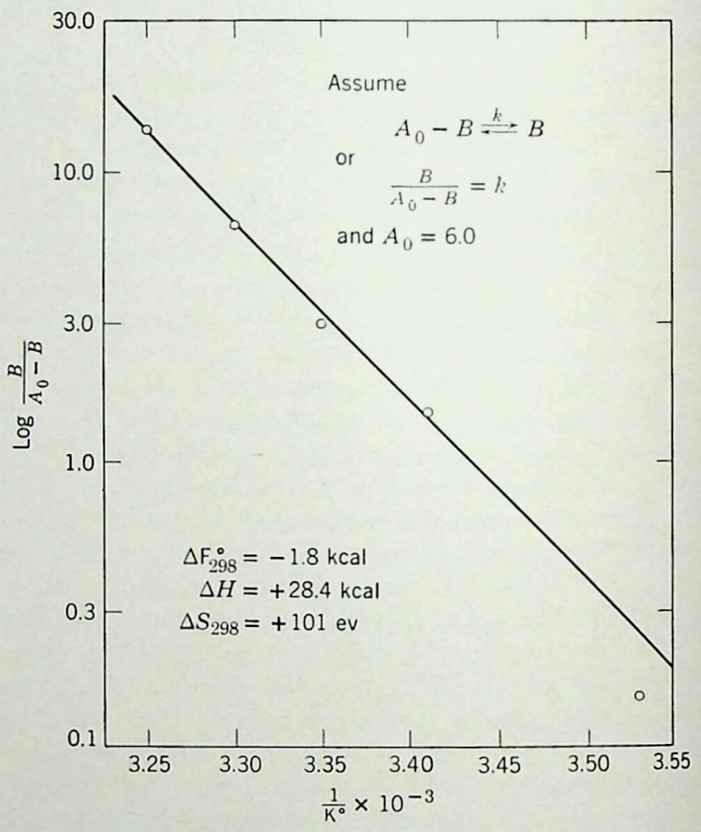


FIG. 8. Log plot of spindle reaction equilibrium *vs* inverse absolute temperature.

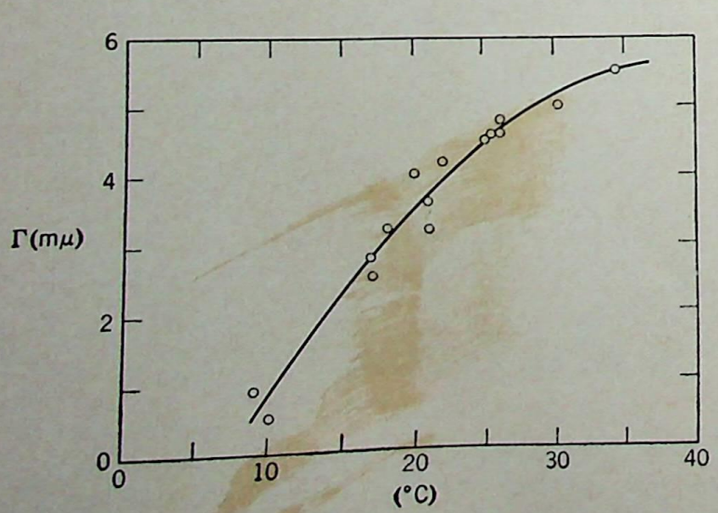


FIG. 7. Equilibrium retardation (Γ) of *Chaetopterus* spindle at various temperatures.

the high heat of reaction and the high positive entropy which nearly cancel each other explain the apparent decrease of entropy (increase of spindle birefringence) at higher temperatures. At high temperature, the still-unoriented protein molecules presumably absorb a considerable amount of heat, thus, for example, releasing bound water which could have prevented their orientation. The melting and randomizing of the bound water then could account for the large increase in entropy (also see Anderson[12]).

It appears that the spindle fibers are regions with high degrees of orientation, although very labile, expressing the orienting influence of the kinetochores and centrioles. As Östergren's earlier observation on the chromosome behavior of a plant *Luzula* also suggests,[30]

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

spindle fibers are undoubtedly almost fluid in nature and are not stably crosslinked gels as the term "fibers" may imply. With dehydrating agents (e.g., alcohol) and in an acidic environment the crosslinking is probably enhanced until the spindle is finally "fixed."

On this basis and from observations described in Sec. 2, anaphase movement of chromosomes may be explained by local reduction in the quantity of oriented material and the consequent shortening of chromosomal fibers. The orienting forces of kinetochore and centriole must be just as actively at work throughout this process. The continuous fibers, either actively elongating or at constant length, could function as supports to counteract the pulling action of the chromosomal fibers. This mechanism of contraction of the chromosomal fibers is similar to that postulated by the author for the action of low concentrations ($<10^{-3}M$) of colchicine.[22] It is, however, in distinct contrast to the mechanism suggested by Swann[47,48] whose microscope lacked the power of resolution for detecting individual spindle fibers.[49]

5. RELATION OF CILIA BEAT AND MITOSIS TO MUSCLE CONTRACTION

The sole evidence for molecular contraction in cilia appears to lie in the x-ray diffraction data on isolated bacterial flagella (Sec. 1). Hypotheses involving mechanisms other than contraction (e.g., differential swelling) also have been postulated, but discriminating experiments are lacking.

Anaphase movement of chromosomes was explained by an orientation equilibrium of spindle fibers (Sec. 4). The action of centrioles and kinetochores—the center of foci of orientation in the spindle—appears similar to that of basal granules of cilia during fibrogenesis (Sec. 3).

Occasionally, cilia are resorbed or re-formed, but the process is much slower than the formation and disappearance of the mitotic spindle at each cell division. It appears that the cilia fibrils are quite stable while the molecules in spindle fibers probably are barely crosslinked (Sec. 4). In comparison, the contractile material in muscle may have a stability lying in between that of cilia and spindle fibers. The primary function of muscle and cilia is repeated rapid contraction, while with the spindle it is a single successful partition of the chromosomes into two new cells.

Mechanisms of muscle contraction are discussed in other papers of this symposium. It is interesting that one of the most widely discussed (and rather widely accepted) current hypotheses is that involving the creeping of two sets of filaments past each other.[50,51]

Regardless of the exact mechanism of choice, one may not overlook the muscle model systems which can contract and produce the same force per cross section as live muscles. This is true in muscle fibers extracted with 50% chilled glycerol, and in an oriented gel fiber formed by mixing two purified muscle proteins, actin and myosin. In either case, contraction is induced specifically by the addition of ATP (adenosine triphosphate) in the presence of magnesium and potassium ions.[52–54]

Hoffmann-Berling has shown further that motility can be induced in glycerol-extracted cells other than muscle, again by the addition of ATP. Thus, he was able to induce glycerinated sperm-tail flagella to undergo prolonged beating, chromosomes to separate, and extracted dividing cells to complete formation of their cleavage furrow (motion picture commercially available[54,55]). These models respond to approximately the same concentration of ATP as muscle models.

To what extent the movements induced by ATP in various cells reflect the same molecular mechanisms still is not clear. For example, the elongation of the central spindle, apparently responsible for the separation of chromosomes in the cell model, is not prevented by the same poisons to which the muscle and cilia model are very sensitive. Furthermore, although the organic triphosphate ATP (and ITP) specifically induces movements in extracted cells, the responding proteins show significant difference in their amino-acid compositions (see Secs. 1 and 2). It is nevertheless encouraging that movements closely resembling those found in living cells can be induced by the same reagent in cells from which much of the complex structures and materials have been removed.

In conclusion, evidence for the long-sought molecular folding is still weak, and no unifying molecular mechanism has been found for cilia beat, anaphase chromosome movement, and muscle contraction. Developments in recent structural and physicochemical analyses are, however, encouraging, and with a concerted intelligent approach, one may acquire before too long a much clearer understanding of the mechanisms underlying these cellular movements.

BIBLIOGRAPHY

[1] J. Gray, *Ciliary Movement* (The Macmillan Company, New York, 1928).

[2] M. Hartmann, *Allgemeine Biologie* (Gustav Fischer Verlag, Stuttgart, 1953), fourth edition.

[3] L. V. Heilbrunn, *An Outline of General Physiology* (W. B. Saunders Company, Philadelphia, 1952), third edition.

[4] D. W. Fawcett and K. R. Porter, J. Morphol. **94**, 221 (1954).

[5] B. J. Krijgsman, Arch. Protistenk. **52**, 478 (1925).

[6] J. R. G. Bradfield, Symposia Soc. Exptl. Biol. **9**, 306 (1955).

[7] D. R. Pitelka and C. N. Schooley, J. Morphol. **102**, 199 (1958).

[8] W. T. Astbury, E. Beighton, and C. Weibull, Symposia Soc. Exptl. Biol. **9**, 282 (1955).

[9] W. J. Schmidt, *Protoplasma-Monographien* (Gebrüder Borntraeger, Berlin, 1937), Vol. II.

[10] E. De Robertis, J. Biophys. Biochem. Cytol. **2**, 319 (1956).

[11] F. S. Sjöstrand, J. Cellular Comp. Physiol. **42**, 45 (1953).

[12] N. G. Anderson, Quart. Rev. Biol. **31**, 243 (1956).

[13] J. G. Carlson, Science **124**, 203 (1956).

[14] L. R. Cleveland, J. Protozool. **4**, 230 (1957).

[15] A. Hughes, *The Mitotic Cycle, the Cytoplasm and Nucleus during Interphase and Mitosis* (Academic Press, Inc., New York, 1952).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

[16] H. Ris in *Analysis of Development*, B. H. Willier, P. A. Weiss, and V. Hambürger, editors (W. B. Saunders Company, Philadelphia, 1955), p. 91.
[17] F. Schrader, Biol. Bull. **67**, 519 (1934).
[18] F. Schrader, "A critique of recent hypotheses of mitosis," *Symposium on Cytology* (Michigan State College Press, East Lansing, Michigan, 1951).
[19] F. Schrader, *Mitosis* (Columbia University Press, New York, 1953), second edition.
[20] C. P. Swanson, *Cytology and Cytogenetics* (Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1957).
[21] E. B. Wilson, *The Cell in Development and Heredity* (The Macmillan Company, New York, 1928), third edition.
[22] S. Inoué, Exptl. Cell Research. Suppl. 2, 305 (1952).
[23] S. Inoué, Chromosoma **5**, 487 (1953).
[24] D. Mazia, Symposia Soc. Exptl. Biol. **9**, 335 (1955).
[25] D. Mazia, Advances in Biol. and Med. Phys. **4**, 70 (1956).
[26] D. Mazia and K. Dan, Proc. Natl. Acad. Sci. U. S. **38**, 826 (1952).
[27] J. G. Carlson, Chromosoma **5**, 199 (1952).
[28] R. Chambers in *General Cytology*, E. V. Cowdry, editor (The University of Chicago Press, Chicago, 1924), p. 235.
[29] S. Inoué, Biol. Bull. **103**, 316 (1952).
[30] G. Östergren, Hereditas **35**, 445 (1949).
[31] W. J. Schmidt, Chromosoma **1**, 253 (1939).
[32] K. Bělař, Ergeb. Fortschrift. Zool. **6**, 235 (1926).
[33] A. W. Pollister and P. F. Pollister, Ann. N. Y. Acad. Sci. **45**, 1 (1943).
[34] M. Bessis and J. Breton-Gorius, Bull. microscopic appl. **7**, 54 (1957).
[35] E. de Harven and W. Bernhard, Z. Zellforsch. u. mikroskop. Anat. **45**, 378 (1956).
[36] J. Rhodin and T. Dalhamn, Z. Zellforsch. u. mikroskop. Anat. **44**, 345 (1956).
[37] Ch. Rouiller and E. Fauré-Fremiet, J. Ultrastructure Research **1**, 289 (1958).
[38] M. H. Burgos and D. W. Fawcett, J. Biophys. Biochem. Cytol. **2**, 223 (1956).
[39] D. W. Fawcett, Intern. Rev. Cytol. **7**, 195 (1958).
[40] F. Meves, Verhandal. Anat. Ges. (Jena) Halle, 152 (1902).
[41] P. R. Gross, Trans. N. Y. Acad. Sci. **20**, 154 (1957).
[42] K. R. Porter in *Harvey Lectures* (Academic Press, Inc., New York, 1957), Vol. LI, p. 175.
[43] J. C. Dan, Physiol. Zool. **21**, 191 (1948).
[44] S. Inoué and K. Dan, J. Morphol. **89**, 423 (1951).
[45] M. M. Swann, Quart. J. Microscop. Sci. **94**, 369 (1953).
[46] Thermodynamical analysis of the data was carried out in 1958 by M. Morales of Dartmouth College Medical School. The author is deeply indebted for his contribution.
[47] M. M. Swann, J. Exptl. Biol. **28**, 417 (1951).
[48] M. M. Swann, J. Exptl. Biol. **28**, 434 (1951).
[49] A. F. Hughes and M. M. Swann, J. Exptl. Biol. **25**, 45 (1948).
[50] J. Hanson and H. E. Huxley, Symposia Soc. Exptl Biol. **9**, 228 (1955).
[51] A. F. Huxley, Progr. in Biophys. and Biophys. Chem. **7**, 255 (1957).
[52] T. Hayashi and R. Rosenbluth, J. Cellular Comp. Physiol. **40**, 495 (1952).
[53] A. Szent-Györgyi, *Chemistry of Muscular Contraction* (Academic Press, Inc., New York, 1951), second edition.
[54] H. H. Weber, *The Motility of Muscle and Cells* (Harvard University Press, Cambridge, 1958).
[55] H. Hoffmann-Berling, Fortschr. Zool. **11**, 142 (1958).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

REVIEWS OF MODERN PHYSICS VOLUME 31, NUMBER 2 APRIL, 1959

45
Fibrous Proteins of Muscle*

ALAN J. HODGE

Department of Biology, Massachusetts Institute of Technology, Cambridge 39, Massachusetts

A FULLY coherent and integrated picture of the structure of muscle in terms of its various components and the changing relationships accompanying or responsible for contraction is still a long way off, as becomes evident later. An adequate description of muscle contraction requires knowledge sufficient to answer questions such as the following: First, which are the components of the myofibril actually participating in the contraction? Second, how does the structure of the individual components and their interrelationships one with another change during shortening? Third, how is the mechanical work produced at the expense of free energy derived from the hydrolysis of adenosine triphosphate (ATP)? The first of these questions is, at present, the only one which can be answered with any certainty. Although many models have been proposed for the contraction process, the physical changes accompanying contraction remain a subject for speculation, largely because of the difficulty in applying appropriate physical methods, such as x-ray diffraction and optical rotation, to the problem. The electron microscope has been employed with striking success in elucidating the details of muscle fine structure but, because of certain inherent limitations, it has been much less useful in following the structural changes accompanying contraction. A final solution, it would seem, can be accomplished only by a careful and rigorous synthesis of data obtained from light and electron microscopy, x-ray diffraction, optical rotation, and other physical and chemical techniques.

Before passing on to a consideration of the individual muscle proteins, it seems appropriate to review briefly some aspects of the progress which has been made since Engelhardt and Ljubimova[1] demonstrated the ATPase activity of "myosin." Later, Szent-Györgyi and his collaborators showed that myosin is a complex of two proteins, actin and L-myosin, and that the association of these components is influenced by the presence of ATP.[2,3] This discovery led to an extended period in which the mechanical and other properties of actomyosin threads were intensively studied. The discovery and development of the glycerine-extracted model system provided further stimulation for the study of artificial model systems in general. Further progress in narrowing the gap between intact muscle and the various extracted actomyosin systems was made by the discovery of substances able to influence contraction in one way or another. Of these, perhaps the best-known is the "relaxing factor" of Marsh.[4] The more-refined physicochemical methods developed in recent years have been applied to advantage in characterizing the various muscle proteins, but they have been relatively unsuccessful in elucidating the molecular changes accompanying contraction phenomena. Furthermore, it is difficult to extrapolate from the behavior of macromolecules in solution to their behavior in a complex paracrystalline multicomponent system like the myofibril. There exists an urgent need for the development of techniques which will allow the determination of configurational changes both at the molecular and macromolecular levels during contraction in intact muscle. X-ray diffraction is the only method currently applicable to this problem, and its application is fraught with very great technical difficulties, stemming largely from the short duration of a twitch and from the very high source intensities required for an adequate diffraction record at low angles. The application of optical-rotation methods appears to be an attractive possibility provided that the physical difficulties arising from the small dimensions and high degree of orientation of the contractile components can be overcome.

Following the pioneering work of Astbury in classifying the fibrous proteins into the α class (kmef group, now known to possess the α-helical configuration[5–7]), and other classes which need not be considered here, Bear[8] showed that the x-ray diffraction patterns of various types of muscle fall into two distinct groups as judged by their low-angle diffraction patterns: (1) the Type I or paramyosin pattern corresponding to an axial repeat of 725 A found only in certain invertebrate muscles possessing the so-called "catch mechanism"; and (2) the Type II diffractions corresponding to axial spacing of about 400 A which are found in all muscles. As becomes evident later, this classification based on x-ray spacings is probably far more reliable as a criterion than some of the biochemical properties such as solubility and amino-acid composition, which have been employed in recent years as bases for nomenclature of the muscle proteins. At first, it was thought that this 400-A periodicity arose from a single component. However, as is seen, x-ray diffraction studies on purified components of striated muscle indicate that most, if not all of them, exhibit periodicities of about 400 A. Furthermore, the macromolecules of many of these Type II components have lengths of about this magnitude, and electron micrographs of myofibrils frequently show an axial spacing of similar dimension. It seems very likely, therefore, that, in the myofibril, the various

* This paper contains descriptions of original work by the author (on tropomyosin and paramyosin) which was aided by a research grant, E-1469 (C1), from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Public Health Service, U. S. Department of Health, Education, and Welfare.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

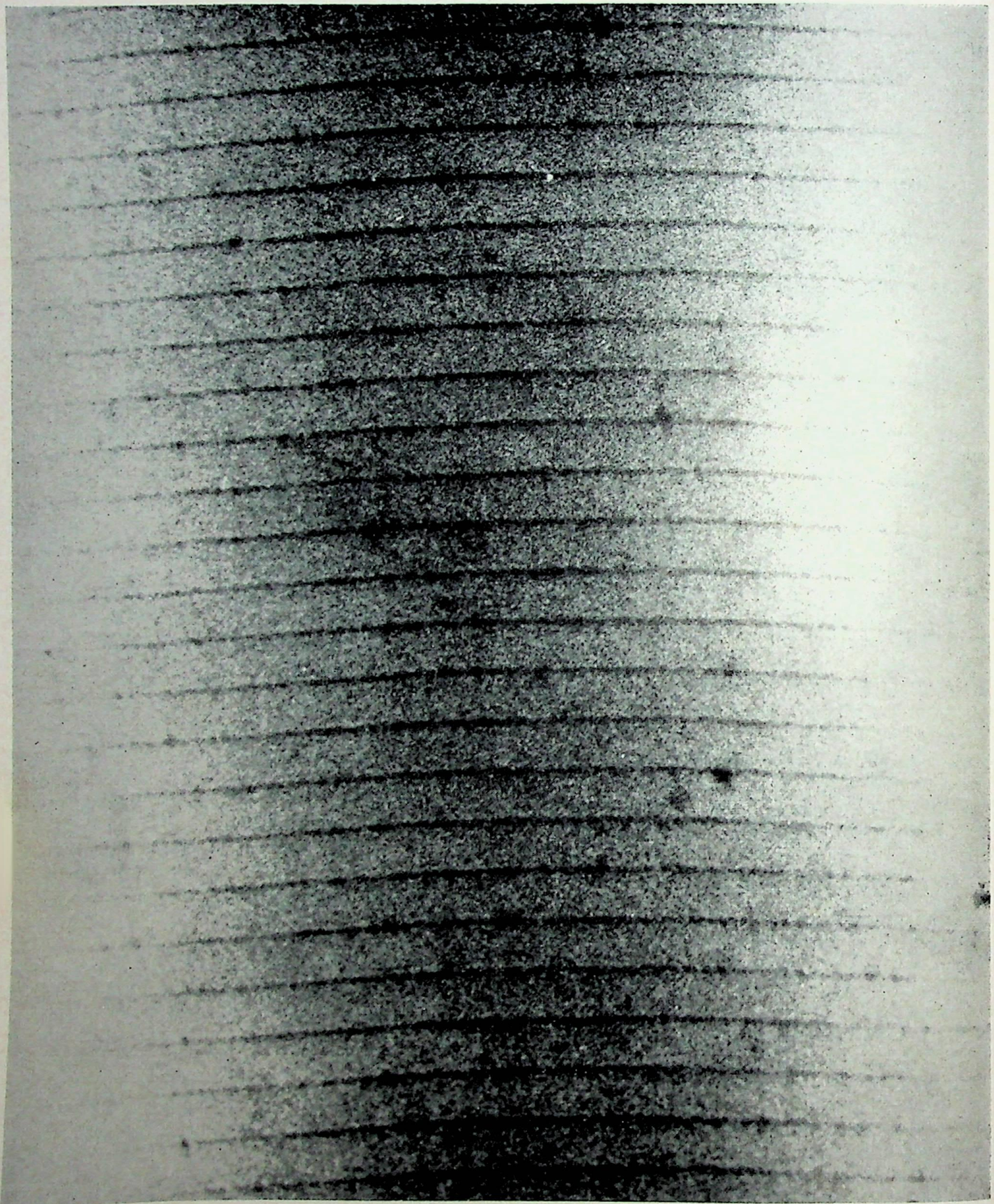


FIG. 1. Crystal of LMM showing the sharp axial banding observed in the electron microscope in the absence of an "electron stain." The very dense transverse striations spaced about 420 A apart are probably owing to binding of salts by the end regions of the LMM molecules [from D. E. Philpott and A. G. Szent-Györgyi, Biochim. et Biophys. Acta **15**, 165 (1954)]). ×200 000.

components interact with one another in orderly fashion by means of reactive sites repeating at distances of about 400 A along the fiber axis.

The results of Huxley and Hanson[9] indicate that intermolecular rearrangement is an important part of the mechanism of contraction in striated muscle. How-

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

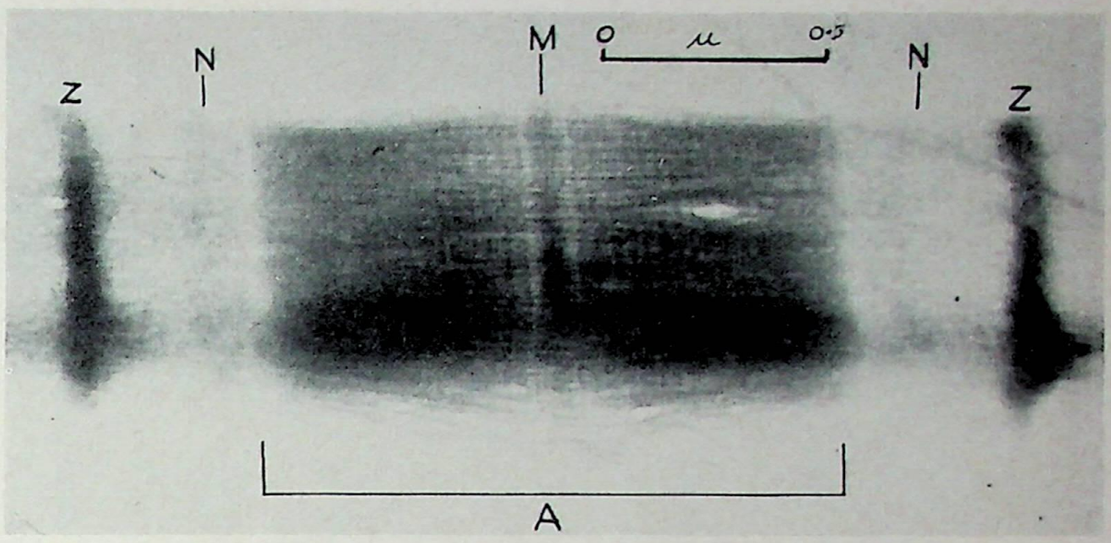


FIG. 2. Electron micrograph of a myofibril isolated from formalin-fixed toad muscle and stained with phosphomolybdic acid, showing the regular 400-A cross-striation in the form of fine, transversely oriented, dense lines [from M. H. Draper and A. J. Hodge, Australian J. Exptl. Biol. Med. Soc. 27, 465 (1949)]. ×64 000.

ever, the results obtained so far do not rule out the possibility that configurational changes are involved as well—the results of detailed correlative x-ray diffraction and other measurements must be awaited before any definite conclusions can be drawn, and in any case, the results obtained with striated muscle are not necessarily applicable to all types of contractile systems.

ACTIN

This component of the actomyosin complex was isolated by Straub.[10] It appears to be a necessary component of the contractile system in muscle and comprises 15 to 20% of the structural proteins in rabbit striated muscle. An outstanding characteristic of actin is its ability to undergo polymerization (the $G-F$ transformation) to form well-defined fibrous elements, a process accompanied by the dephosphorylation of strongly bound ATP. The monomeric form of actin is one of the most difficult muscle proteins to characterize from a physicochemical point of view since the presence of salt favors the formation of the fibrous form. However, it seems likely that the monomeric unit has a molecular weight of about 60 000 (Table I, from work by Cohen and A. G. Szent-Györgyi[11]) with physical dimensions rather more uncertainly defined. The work of Astbury[22,23] and of Selby and Bear[24] indicates that actin exhibits an axial-repeat period of about 400 A in conformity with the other proteins of vertebrate striated muscle and of smooth muscle, both vertebrate and invertebrate. Actin has a pronounced tendency for complex formation with myosin, but is devoid of enzymatic properties.

TABLE I. Dimensions of fibrous-muscle proteins [from C. Cohen and A. G. Szent-Györgyi, paper read at International Biochemistry Congress, Vienna (1958)].

	M.W.	Est. length (A)	Est. axial ratio
Actin	57 000[12]–74 000[13]	290[13] ?	12[13] ?
Myosin	450 000[14–17]	•1600	50
HMM	230 000[18]–330 000[19]	400[18]	15–20[18]
LMM	96 000[18]–140 000[19]	550[18]	30–40[18]
LMM fr. 1	110 000 –120 000		30–40
Tropomyosin	53 000[20]	385[20]	25[20,16]
Paramyosin	134 000[21]	1400[21]	80[20]

MYOSIN

Myosin is the major constituent of most types of muscle, e.g., it accounts for 55 to 60% of the structural proteins of rabbit skeletal muscle. It has ATPase activity which is calcium activated and magnesium inhibited. It is a globulin, precipitating at low ionic strength, and can be extracted from muscle at neutral pH with solutions of ionic strength higher than about 0.6. Characterization of the myosin molecule has been a subject of controversy for many years, but there seems now to be general agreement that the molecule has a length of about 1600 A, an axial ratio of about 50, and a molecular weight of about 420 000 (Table I).

Brief exposure of myosin solutions to proteolytic enzymes such as trypsin,[25–27] chymotrypsin,[28] and subtilizin[29] degrades the myosin molecule into two fairly well-defined components, heavy meromyosin (HMM) and light meromyosin (LMM). The well-known properties of the myosin molecule are divided between these fragments. LMM behaves physically very much as does myosin. It precipitates at low ionic strength and has a molecular weight of about 100 000. However, it has no ATPase activity and does not combine with actin. On the other hand, HMM has all of the ATPase activity of the myosin from which it was derived, combines with actin in the same proportions as does myosin, but differs from myosin in being soluble in solutions of low ionic strength. Reported molecular weights range from 230 000 to 320 000 (Table I). HMM, like myosin, does not form ordered precipitates. LMM, on the other hand, under appropriate conditions, forms highly ordered needle-shaped crystals exhibiting a very striking axial period of about 420 A in the electron microscope, even in the unstained condition, as shown in Fig. 1 from work by Philpott and Szent-Györgyi.[30] It seems very likely that

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

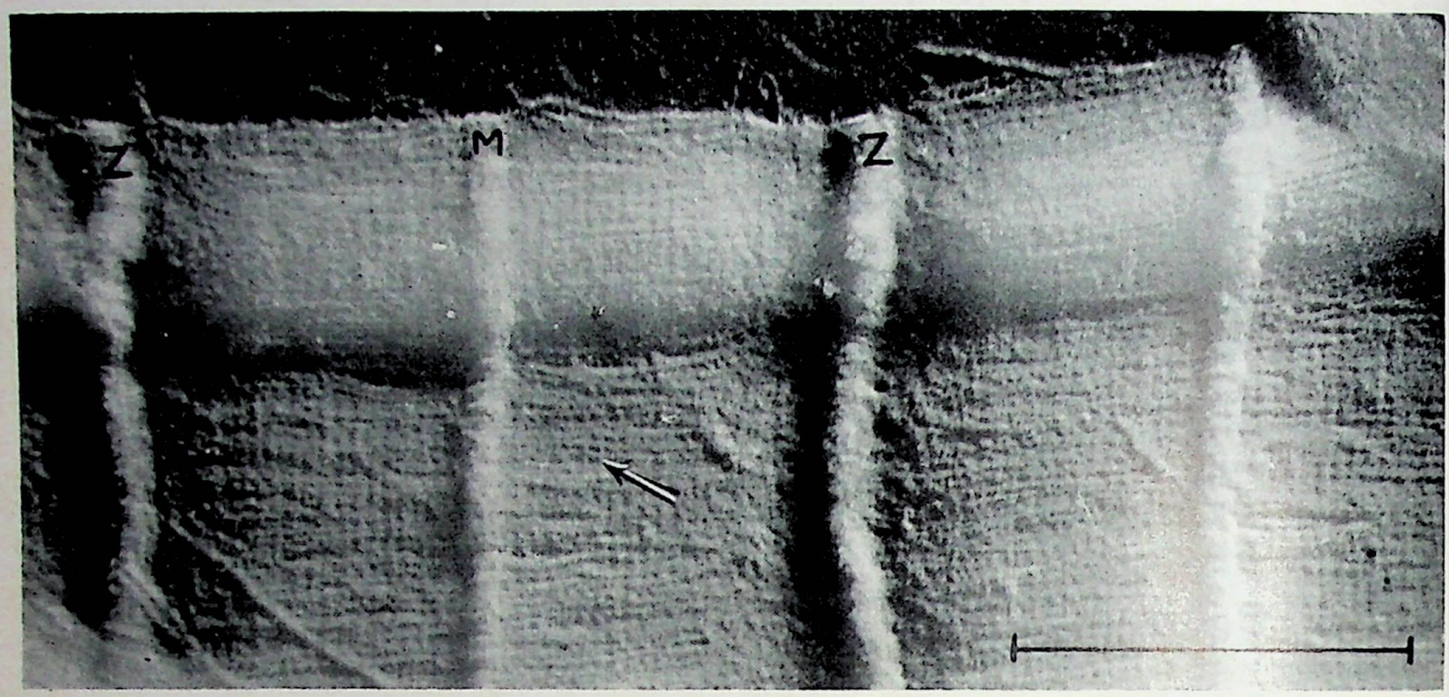


FIG. 3. Semicontracted myofibril of toad muscle, shadowed with platinum, showing the regular axial period apparently associated with the presence of transverse bridges (arrow) [from M. H. Draper and A. J. Hodge, Australian J. Exptl. Biol. Med. Soc. 27, 465 (1949)]. ×45 000.

the fine cross-striations evident in these crystals arise from a selective binding of inorganic ions at specific sites located about 400 A apart (possibly at the ends of the LMM "molecules"). This result is in good agreement with the fact that isolated myofibrils, when shadow-cast or stained with phosphotungstic acid (PTA) or phosphomolybdic acid, exhibit a sharp and regular cross-striation[31,32] (Figs. 2 and 3) with an axial period of about 400 A in the relaxed condition, and it agrees with the evidence obtained by Draper and Hodge[33,34] concerning the distribution of bound mineral in myofibrils after electron-induced microincineration in the electron microscope.† These authors found that the mineral residue from well-washed myofibrils obtained from formalin-fixed muscles is present as fine cross-striations with an axial spacing of about 400 A (Fig. 4), a result suggesting that the various inorganic and organic components of muscle interact one with another by means of reactive sites or groups located about 400 A apart. Further evidence[36] for this view is provided by the regular cross striations observed in thin sections of muscle (Fig. 5). There is in addition, some electron-microscopic evidence to suggest that this fundamental period may shorten during contraction,[37] a result which, if borne out by further investigation, strongly suggests that contraction must involve a configurational change in at least some of the macromolecular components of the myofibril. On the other hand, the sliding filament model, in which contraction is achieved by progressive interdigitation of actin and myosin filaments, does not require such configurational changes, and it should be noted that no changes have so far been observed in the wide-angle x-ray pattern of muscle during contraction. However, such configurational changes

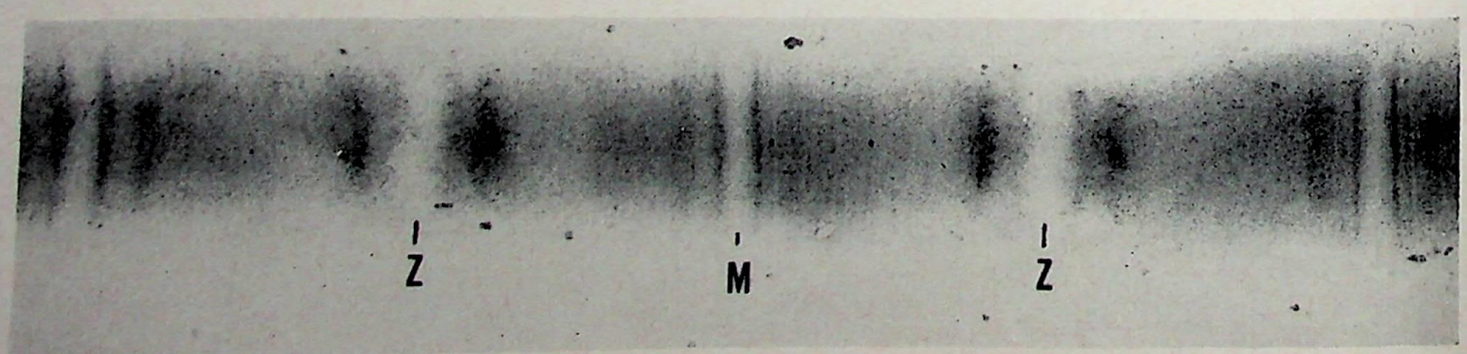


FIG. 4. Formalin-fixed rabbit myofibril after electron-induced microincineration in the electron microscope to illustrate the distribution of the bound mineral residue within the sarcomere in the form of regularly spaced, fine transverse striations spaced about 400 A apart [from M. H. Draper and A. J. Hodge, Nature 163, 576 (1949)]. ×30 000.

† The work of Leisegang[35] provides a rationale for the microincineration of organic material induced by high beam intensities in the electron microscope.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

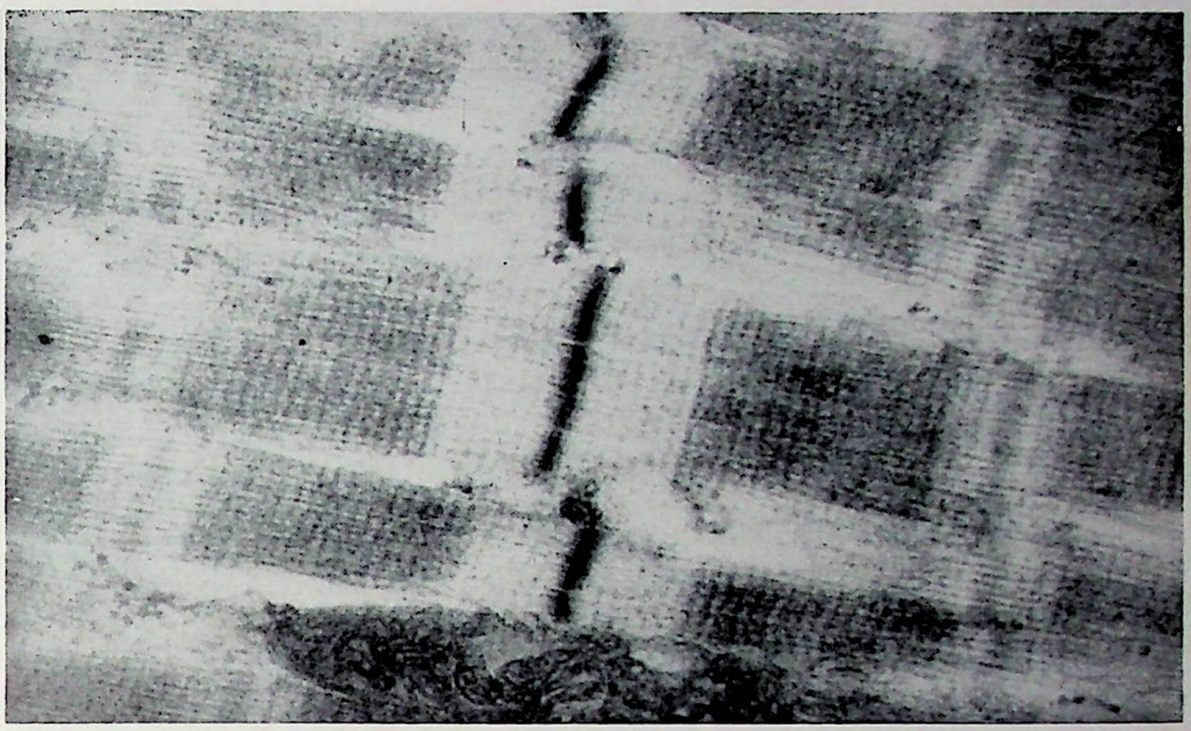


FIG. 5. Thin longitudinal section of rabbit muscle fixed in buffered OsO_4 and stained with PTA, showing the regular cross-striation present in all bands of the sarcomere [from A. J. Hodge, H. E. Huxley, and D. Spiro, J. Exptl. Med. 99, 201 (1954)]. ×40 000.

would be difficult to detect if, as seems possible, a sequential shortening of small contractile units (perhaps 400 A long) is involved in the mechanism of contraction.

A number of recent experimental investigations point toward the rather disquieting possibility that the myosin "molecule" may be an artifact of the extraction methods used to isolate this protein. Thus, although it has so far proved necessary to employ proteolytic enzymes such as trypsin, chymotrypsin, and subtilizin in order to obtain the meromyosins from preparations of myosin, there is as yet no convincing evidence to indicate that the splitting of peptide bonds is necessary for the liberation of the meromyosins from the parent macromolecule. Middlebrook[29] was unable to demonstrate the presence of C-terminal groups in the ratios expected from the known specificities of trypsin and chymotrypsin. Furthermore, the degradation of the meromyosins by these enzymes is a much slower process than their production from the parent macromolecule, a result indicating the presence of a highly sensitive region within the myosin molecule, and the results of turnover-rate studies indicate a metabolic independence of the two meromyosins. Thus, Velick[38] has shown that the turnover rate of phenylalanine is about

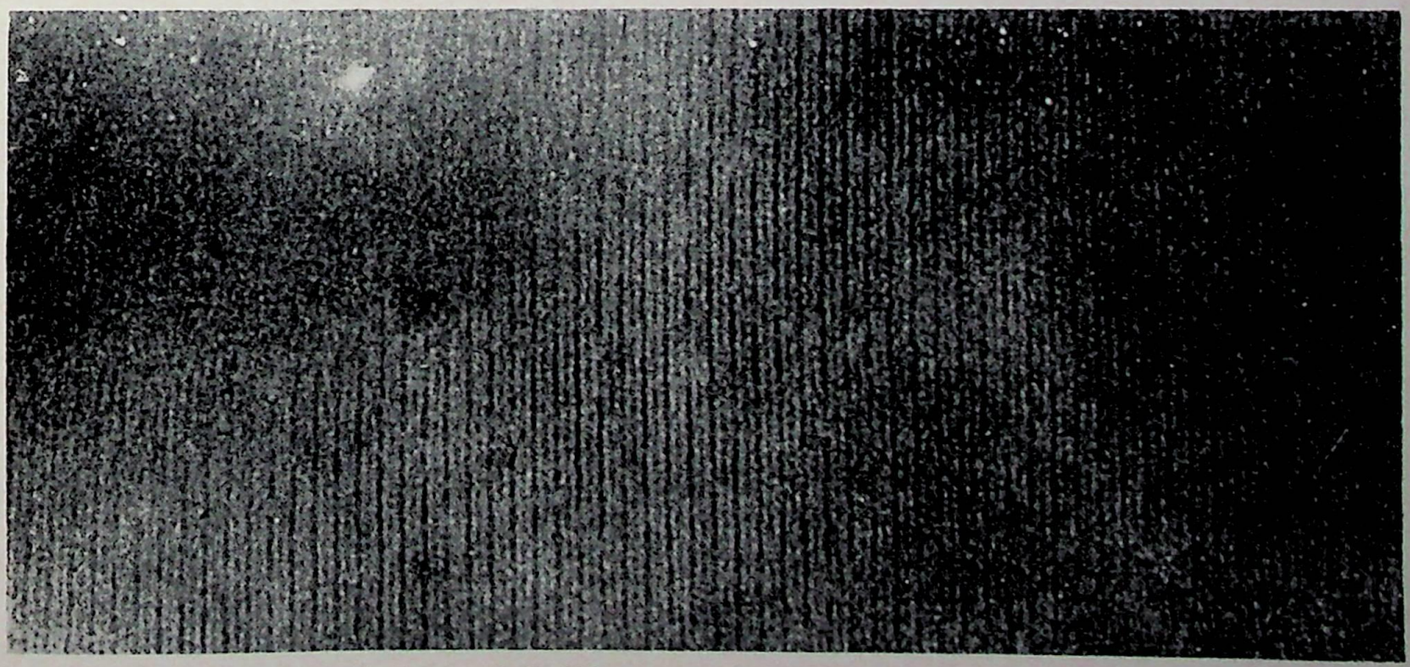


FIG. 6. Electron micrograph of a crystal of rabbit tropomyosin deposited on a supporting film and stained with PTA in *p*H 4.2 phosphate buffer. The main spacing is about 200 A, and a definite intermediate line is present. ×125 000.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

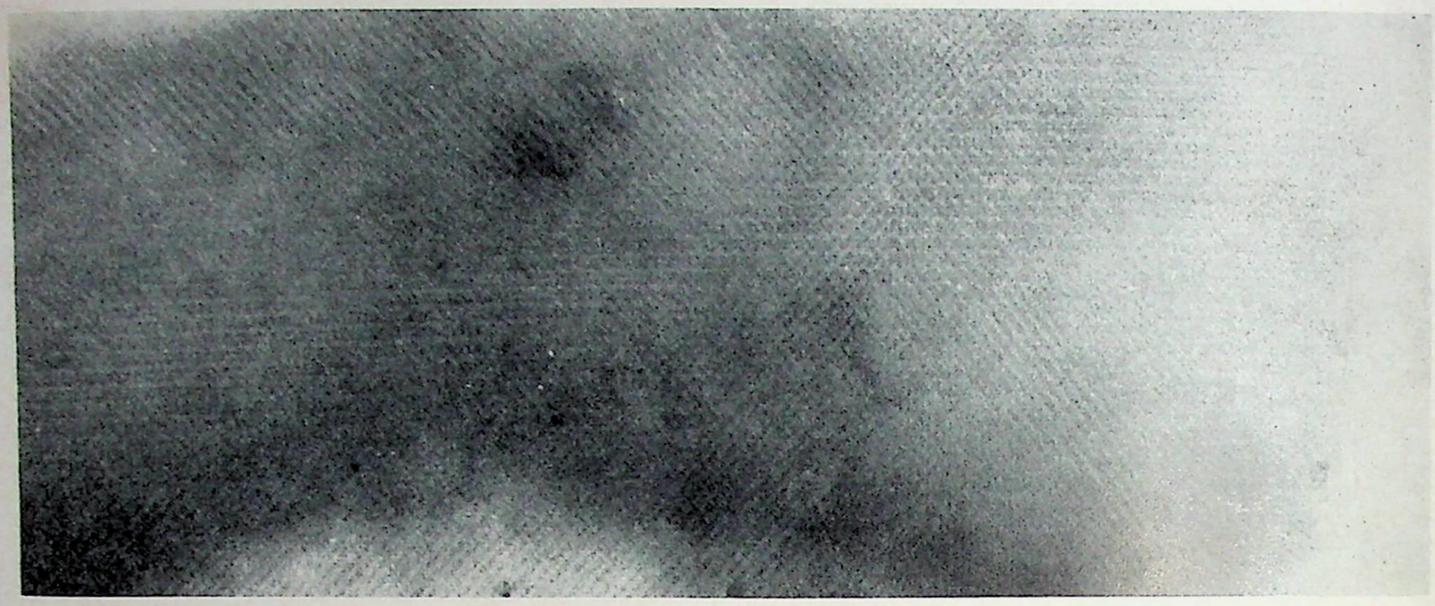


FIG. 7. Crystal of rabbit tropomyosin, stained with PTA, showing the crossed-grid appearance characteristic of many of these crystals in electron microscopic preparations. ×70 000.

five times more rapid in LMM than in HMM. Similarly, Schapira *et al.*[39] have observed that the incorporation of glycine is more rapid for LMM than for HMM. These results, on their face value, may be interpreted as indicating that the meromyosins represent precursors of the myosin macromolecule. However, there is also evidence derived from careful experiments with fluorescent antibodies to suggest that at least a part of the meromyosins in the myofibril are distributed independently of each other and in a highly characteristic pattern. This is discussed later. If nothing else, the sum total of the foregoing results must sow a seed of doubt concerning the almost universally accepted assumption that myosin *per se* is a well-defined macromolecular species which by some relatively simple interaction with other species such as actin is able to bring about the phenomenon of contraction.

TROPOMYOSIN

This protein component, which is universally distributed in small concentrations in a wide variety of muscles, was discovered by Bailey[40] and is remarkable in that it is the first and probably the only fibrous protein thus far obtained in a truly three-dimensional crystalline form. The crystals of tropomyosin are platelike and have an unusually high degree of hydration (80 to 90%). This latter property probably results from the fact that tropomyosin, unlike many of the fibrous proteins of muscle (which tend to form one-dimensional paracrystalline arrays), exhibits a strong tendency for the formation of three-dimensional crossed-grid net-

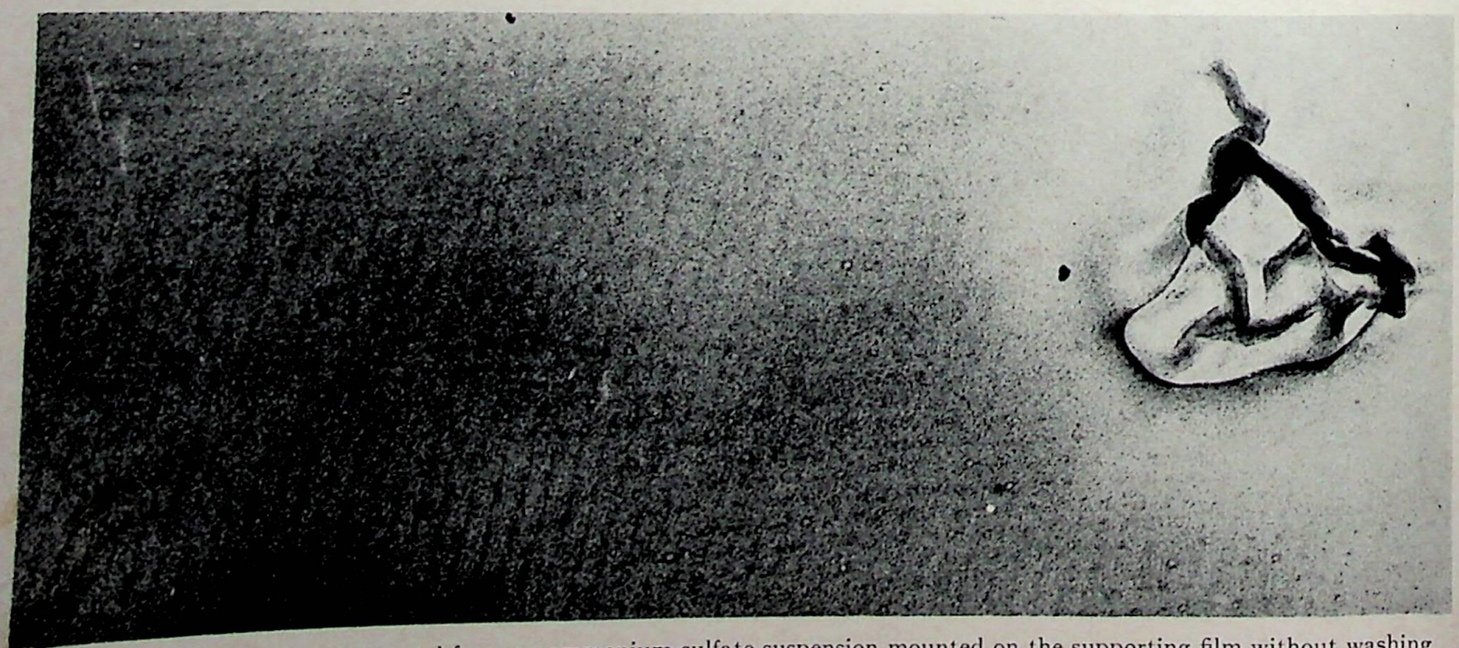


FIG. 8. Rabbit tropomyosin crystal from an ammonium-sulfate suspension mounted on the supporting film without washing or staining. The points and lines of high density probably correspond to regions where salt is accumulated. ×135 000.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

works. A crystal of tropomyosin can, in fact, be regarded as a very highly ordered gel. Tropomyosin has no enzymatic activity and, as its name implies, was at one time thought to be a precursor of myosin itself. However, there is no substantial evidence for this view, and indeed, the role played by tropomyosin in muscle function remains a subject for speculation.

Tropomyosin, the water-soluble component originally defined by Bailey,[40] appears to be present in small amounts in all muscles (e.g., it accounts for about 4 to 5% of the fibrous proteins of rabbit striated muscle) and is present even in those invertebrate muscles that exhibit Type I low-angle diffraction patterns (i.e., those that have paramyosin as well as the actomyosin system). It is of interest to note that the tropomyosin molecule with a molecular weight of about 53 000 (Table I), has a length of about 400 A, as estimated from physicochemical evidence, a result in keeping with the lengths of all of the other components and subunits in the Type II system so far described.

Examination of tropomyosin crystals in the electron microscope in this laboratory has proved instructive and the spacings observed appear compatible with physicochemical estimates of the molecular length. A correlation with low-angle x-ray diffraction is currently being carried out.[41] The results obtained when a preparation of small tropomyosin crystals is placed on a grid, stained with buffered phosphotungstic acid (*p*H 4.2) or allowed to dry without staining are illustrated in Figs. 6–9. A frequently observed pattern is one comprising striations spaced about 200 A apart (often with intraperiod lines) or a crossed-grid network with periods of the same magnitude in two directions (Figs. 6 and 7). In unstained crystals (Fig. 8), the periodic structure appears rather as a pattern of dots, strongly suggesting that inorganic ions are occluded or bound at, or near, the sites of interaction of the long thread-like macromolecules arranged in an ordered gel structure. The square two-dimensional pattern with spacings of about 400 A (Fig. 9) affords yet another example of the remarkable capacity of tropomyosin to form open, ordered network structures, a property which may well have some significance in relation to the transverse bridges known to connect the longitudinal filamentous elements in striated muscle at regular intervals along the fiber axis. The regular outlines of the tropomyosin crystals become evident if they are fixed, embedded, and sectioned by conventional procedures (Figs. 10 and 11). However, dimensional analysis is rather more uncertain than in the cases already mentioned, since it is technically difficult to define the plane of the section in relation to a particular crystallographic plane. Nevertheless, the three-dimensional ordered net structure of the crystals is often clearly discernible (Fig. 12). This type of net structure is reminiscent of that observed in collagenous tissues such as Descemet's membrane[42] and in the pharyngeal structures of certain protozoa.[43] It may represent an important structural principle in ordered biological systems and may not be irrelevant to the process of contractility itself.

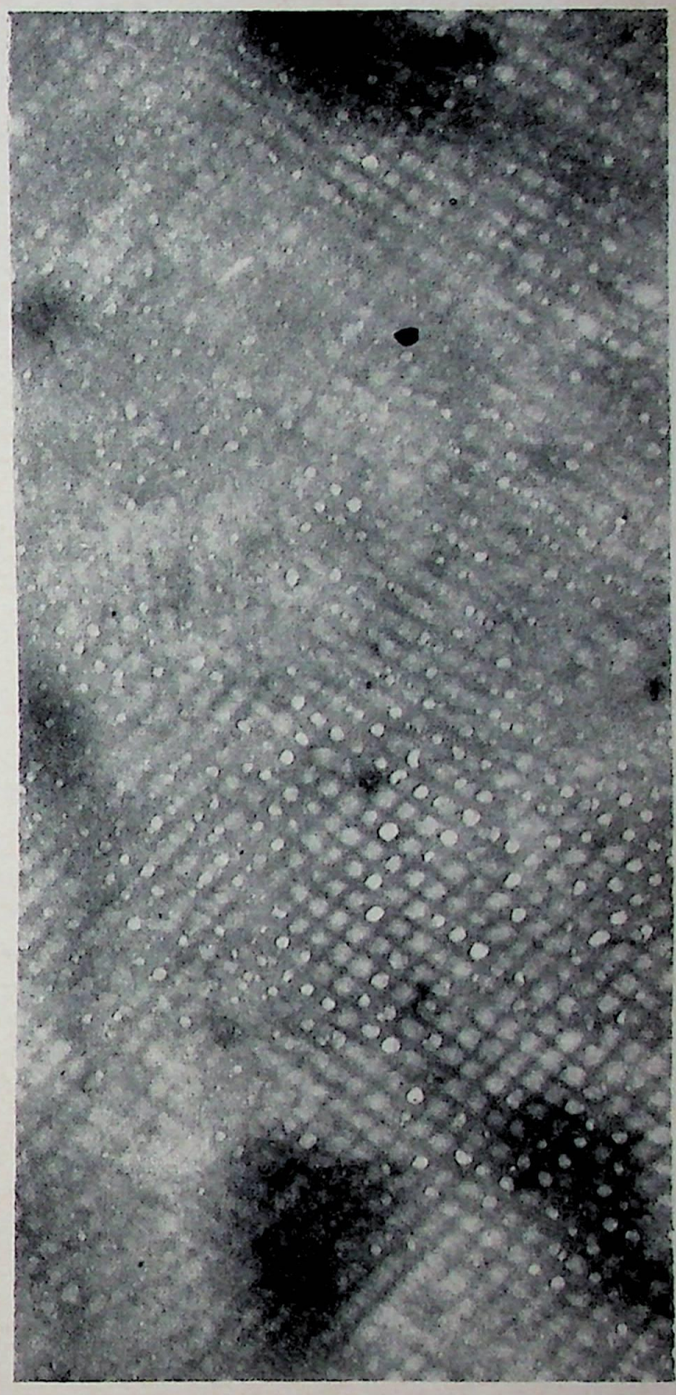


FIG. 9. Open, two-dimensional, net structure observed in a rabbit tropomyosin crystal preparation. The spacing of about 400 A corresponds to the length of the tropomyosin molecule as determined by physicochemical techniques (Table I). ×75 000.

As has been seen, there exists a rationale for an orderly interaction of the muscle components so far described in terms either of their lengths being about 400 A or of there being subunits of this length. It seems clear, therefore, that the total x-ray pattern (both meridional and equatorial) of whole muscle must contain information concerning the precise state of interaction of the macromolecular components for any given state of the contractile system, and offer the

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

FIG. 10. Electron micrograph of a thin section of rabbit tropomyosin crystals, fixed in buffered, osmium-tetroxide solution, stained with PTA, and embedded in *n*-butyl methacrylate. Note the regular outlines of the crystals and the precise cross-striation corresponding to a spacing of about 200 A. ×20 000.

possibility of following the physical changes accompanying contraction. However, as already noted, there are severe technical problems to be overcome in any such investigation.

PARAMYOSIN

The term paramyosin was introduced by Hall *et al.*[44] to describe a major component of certain specialized molluscan and annelid muscles (Table II),[45] which possess the ability to maintain a rigor-like contracture for long periods of time and which were described in the older literature as having a "catch mechanism." It is this protein that is predominantly responsible for the Type I low-angle diffraction pattern obtained by Bear[8] from these muscles, a highly characteristic pattern comprising reflections corresponding to a fundamental repeat period of 725 A, with every fifth reflection highly accentuated. The pattern can be interpreted formally (see Bear and Selby[46]) in terms of a net structure of the type shown in Fig. 13. "Paramyosin fibrils," which are easily isolated in native form from the adductor muscles of certain marine mollusks such as the clam, *Venus mercenaria*, exhibit a characteristic band structure with a period of 145 A in the electron microscope, together with a spot pattern, the symmetry of which is such that the true repeat period of this composite pattern is 5×145 A=725 A (Fig. 13). Since this spot

TABLE II. Relative intensities of the myosin and paramyosin x-ray diffraction systems exhibited by various muscles [from F. O. Schmitt, R. S. Bear, G. E. Hall, and M. A. Jakus, Ann. New York Acad. Sci. **47**, 799 (1947)].[a]

Muscle	Myosin diffractions	Paramyosin diffractions
Mytilus adductor	+	+++++
Venus adductor, w	++	++++
Anodonta adductor, w	++	+++
Anodonta adductor, t	+++	+++
Mya adductor	+++	+++
Venus adductor, t	++++	++
Pecten adductor, w	++	++
Phascolosoma retractor	+	+
Dog retractor penis	+	...
Pecten adductor, t	+	...
Thyone retractor	+	...
Frog sartorius	+	...

[a] Rough visual estimates of the relative intensities of the myosin and paramyosin diffractions on the patterns of the various muscles are indicated by +; w and t refer, respectively, to the "white" and "tinted" components of the muscles which possess them.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

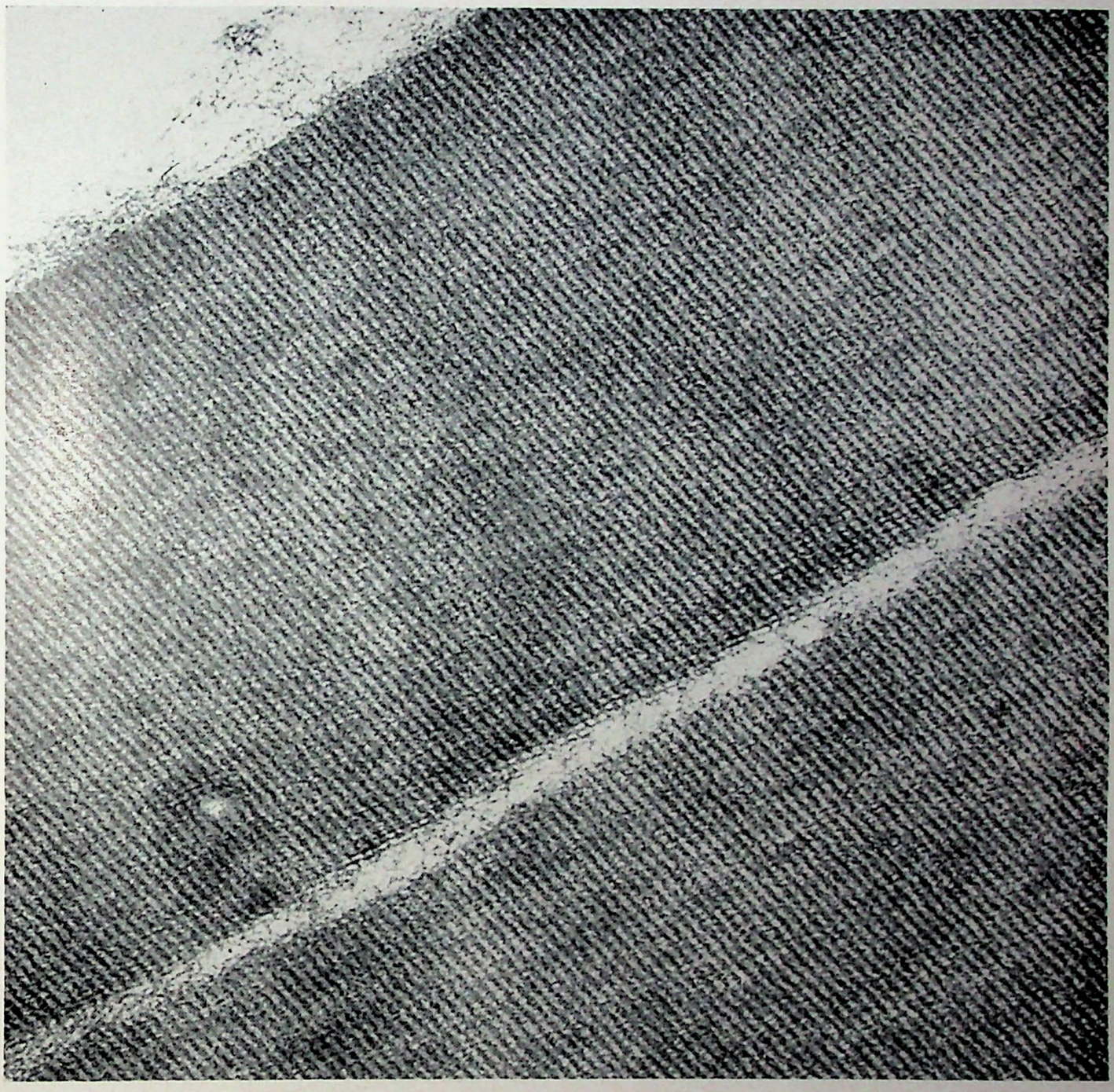


FIG. 11. Higher magnification view of two tropomyosin crystals in the same preparation as Fig. 10, showing the regular open-network lattice characteristic of these crystals. ×95 000.

pattern is frequently absent in these native isolated fibrils, and because the major component of these fibrils can be extracted and reconstituted into fibrils in which no spots are visible but which exhibit band patterns with fundamental repeat periods of 725 A (Fig. 16), it seems likely that the spots reflect the presence of one or more additional components within the fibrils. In any case, it is proposed here to restrict the term paramyosin to the major component of these native fibrils. A composite structure for the native paramyosin fibrils is further indicated by some recent work[47] in which transverse sections of adductor muscle show that the fibrils are built up of a number of thin layers.

In 1952, the author[48,49] found that paramyosin could be quantitatively precipitated from acid solutions in the form of ordered tactoidal fibrils with an axial period of about 1400 A and with symmetrical intraperiod band structure, (Fig. 14) rather than the "polarized" band structure characteristic of the native 725-A period. At that time, Schmitt *et al.*[50] were in the process of analyzing the band structures of the newly discovered segment long-spacing (SLS) and fibrous long-spacing (FLS) forms derived from soluble collagen (asymmetric and symmetric band patterns, respectively), and had deduced that the collagen macromolecule was probably four times the length of the axial period (*ca* 700 A)

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

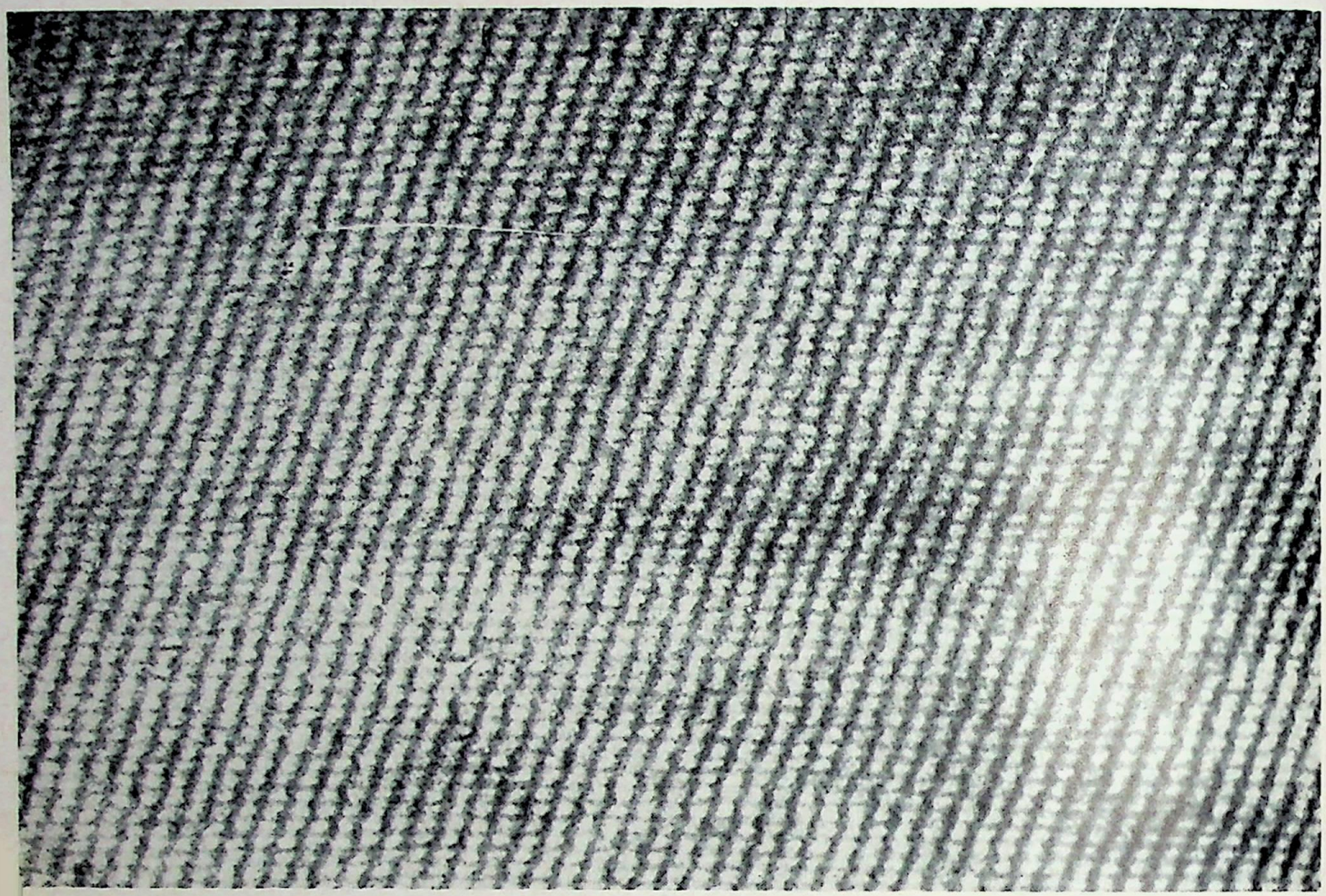


FIG. 12. Thin section of a rabbit tropomyosin crystal showing the characteristic open-network lattice. This type of structure probably accounts for the very high degree of hydration of the tropomyosin crystals. ×190 000.

displayed by native-collagen fibrils. According to their concept, the SLS forms arose by a parallel packing of the macromolecules with like ends in register, the FLS by an antiparallel packing, thus accounting for the symmetrical band structure of FLS. Consequently, by analogy with the collagen picture, it seemed likely that the macromolecules of paramyosin were about 1400 A long, and that they were packing in antiparallel array to form an FLS-type banded structure. This prediction was borne out[48,49] by the results of sedimentation, diffusion, viscosity, and light-scattering investigations, which were necessarily rather crude because of the low ionic strengths required to keep the protein in solution under these acid conditions. However, the data were sufficiently good to indicate the presence in the solutions of particles with lengths between 1200 and 1500 A and with axial ratios of about 70. Thus, the combined electron-microscopical and physicochemical evidence established the length of the paramyosin macromolecule (or at least of the kinetic unit in solution) with some certainty as being about 1400 A. The more recent and more accurate measurements of Kay,[21] carried out on paramyosin in neutral solution at high ionic strength, appear to confirm this value for the length of the paramyosin macromolecule.

Bailey[51] obtained paramyosin in purified form, and, on the basis of similarities in the amino-acid compositions of tropomyosin and paramyosin, was led to call the latter "insoluble tropomyosin."[52] Similarly, Kominz *et al.*[53] refer to paramyosin as "tropomyosin A" to distinguish it from "tropomyosin B" (i.e., the original tropomyosin of Bailey). This profusion of nomenclature seems unwarranted, and it would seem advisable at the present time to retain the term paramyosin, since there exist definite differences in solubility properties, crystalline form, and x-ray and electron-microscopic periodicities which clearly differentiate this protein from the original water-soluble tropomyosins of Bailey. Paramyosin has no enzymatic activity, according to Szent-Györgyi,[54] does not combine with actin, and there is as yet no evidence to indicate a possible function for it in relation to the specialized properties of the "catch" muscles.

A remarkable variety of fibrous structures can be formed from paramyosin solutions under various conditions of *p*H and ionic strength. Those so far obtained include fibrils with axial periods of *ca* 70, 145, 725, and 1800 A. (This last, together with one of 360 A, was first observed by Locker and Schmitt.[55]) In the case of collagen, it can be shown that the native type, 700-A

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

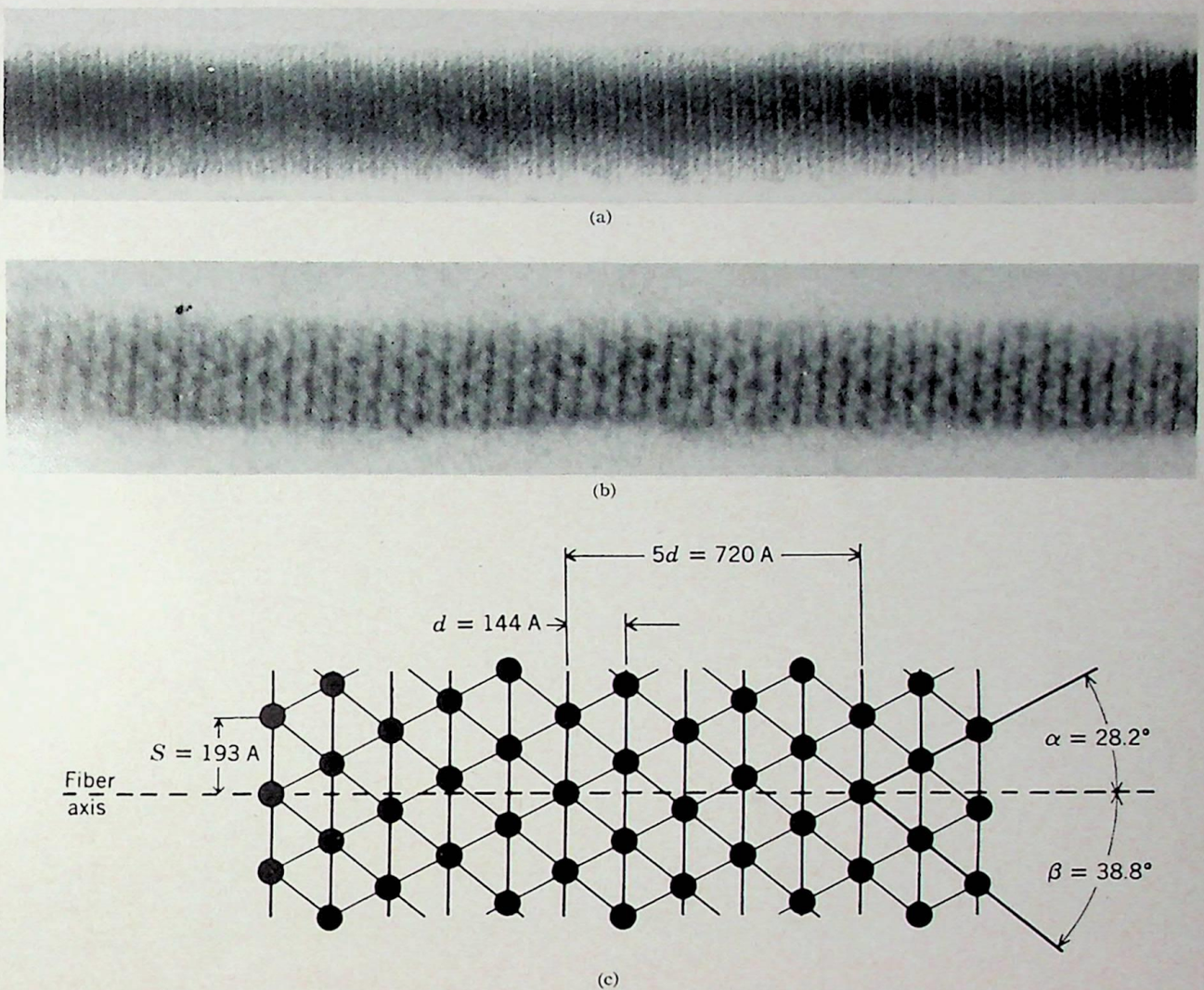


FIG. 13. Native "paramyosin fibrils" isolated from the adductor muscles of *Venus mercenaria* in dilute salt solutions. These usually show (a) a transverse cross-striation with an apparent period of 145 A,[48,49] (b) a pattern consisting of the 145-A cross-striation and a superposed dot pattern of symmetry such that the true repeat period is 5×145 A=725 A [from F. O. Schmitt, R. S. Bear, C. E. Hall, and M. A. Jakus, Ann. N. Y. Acad. Sci. **47**, 799 (1947)]. (c) shows the node pattern derived from electron micrographs. This formal net structure is in good agreement with the results of low-angle x-ray diffraction studies [from C. E. Hall, M. A. Jakus, and F. O. Schmitt, J. Appl. Phys. **16**, 459 (1945)].[8] (a) and (b) ×190 000.

periodicity can be obtained by a parallel packing in which the tropocollagen macromolecules (2800 A long) are staggered with respect to one another by one-quarter of their length. The key to such a synthesis of band structure is the availability of the SLS band structure, which is in effect a "fingerprint" of the tropocollagen macromolecule. Attempts to produce segment-type structures from paramyosin solutions have so far proved unsuccessful, but it seems likely that parallel, antiparallel, and staggered arrangements involving axial displacements of one-tenth the molecular length (i.e., 1400 A/10) or integral multiples thereof could explain the various band patterns encountered thus far (Figs. 13–18). A feature of interest in relation to these band patterns is the frequent occurrence of transitions from one type of packing to another. Figure 18 clearly shows such a transition from the 1800-A type to the 145-A spacing, Fig. 17 a rhythmic transition from the 145-A spacing to a *ca* 70-A spacing and vice versa.

X-RAY DIFFRACTION AND ROTATORY DISPERSION

All of the proteins considered up to this point, with the exception of actin, give wide-angle x-ray diffraction patterns in which the prominent feature is the 5.1-A meridional reflection, indicating that at least a significant proportion of the polypeptide chains possess the α-helical configuration observed by Pauling and Corey.[5,6] This is true also of the meromyosins,[11] LMM giving a good pattern, with HMM being more variable since it is rather easily denatured. However, the x-ray diffraction method does not lend itself to estimation of the noncrystalline but helical parts of the molecule. The helix content of proteins in aqueous solution can best be estimated at the present time by rotatory-dispersion measurements, and the equation by Moffitt[56] describing the rotatory dispersion of the α-helix can be used for this purpose with the muscle proteins, subject to certain theoretical limitations, and on the assumption that the

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

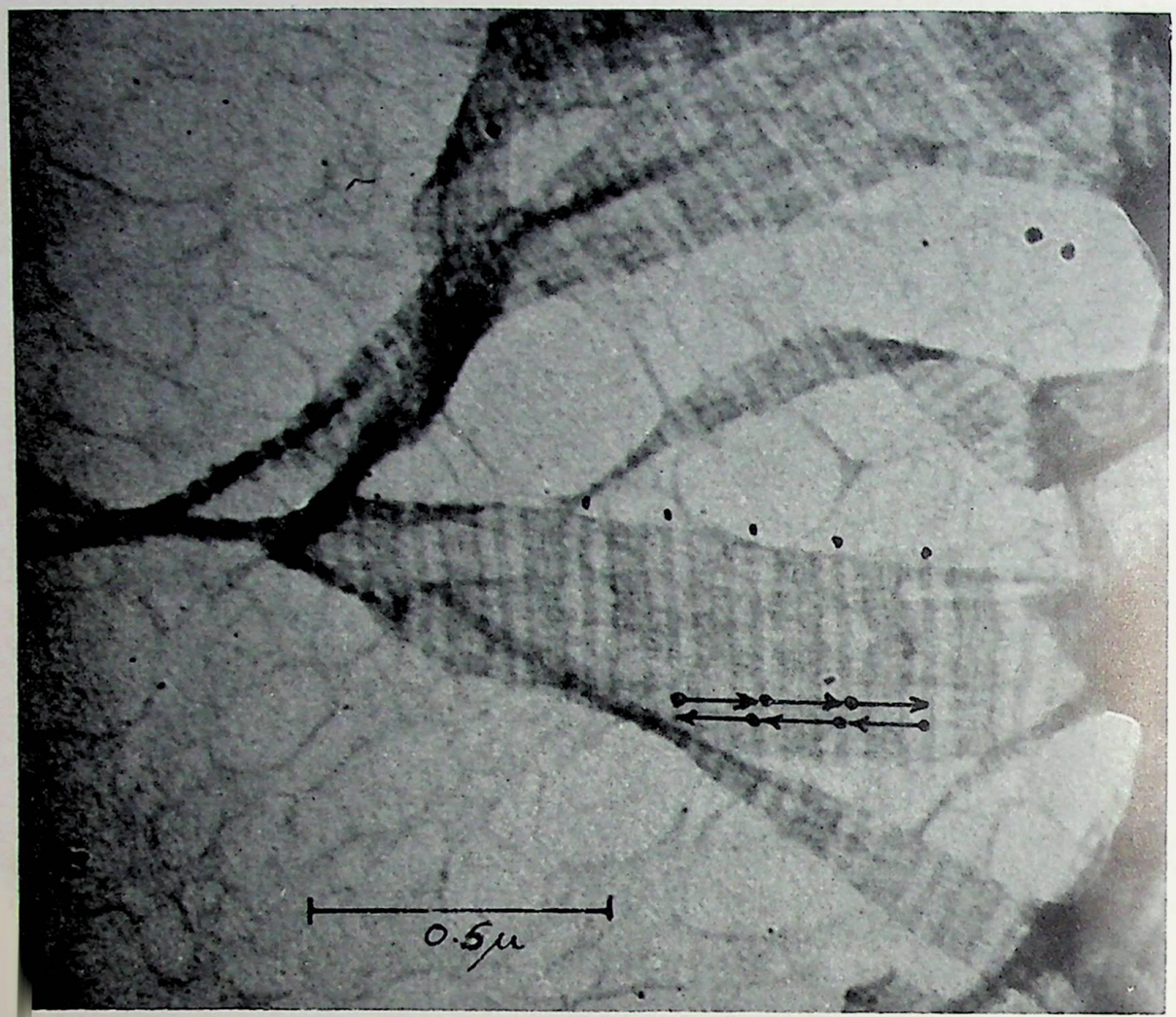


FIG. 14. Fibrils reconstituted from an acid solution of paramyosin showing an axial repeat period of about 1400 A. Note that the band structure within each repeat period is arranged symmetrically. This is interpreted as arising from an antiparallel packing of rod-like molecules about 1400 A long [from A. J. Hodge, Proc. Natl. Acad. Sci. U. S. 38, 850 (1952)]. ×70 000.

entire helical content is in the right-handed α-helical configuration. Table III, from the work of Cohen and Szent-Györgyi,[11] shows the estimated α-helix content in aqueous solutions of the muscle proteins and fragments under discussion. It can be seen from this table also that the extent of the helical regions seems to be dependent upon the proline content. The proline contents of the highly helical muscle components (LMM fr. I, tropomyosin, and paramyosin) are very low, while those of myosin and HMM (which exhibit the lowest helix content) are correspondingly high. On a statistical basis, the data indicate that each proline residue prevents about 20 other residues from participating in α-helix formation.

TABLE III. Helix content and proline and cystine concentration of muscle proteins [from C. Cohen and A. G. Szent-Györgyi, paper read at International Biochemistry Congress, Vienna (1958)].

Protein	Wt. % helix	Equ. cystine in 10^5 g	Wt. % proline	No. of nonhelical residues per proline residue
LMM fr. 1	100	0	0.22	
Tropomyosin	94	1.65	0.35	15
Paramyosin	91	0	0.21	36
LMM	74	0.65	0.97	23
Myosin	56	0.6	2.08	18
HMM	45	1.2	2.87	16

LOCALIZATION OF THE MUSCLE PROTEINS IN THE MYOFIBRIL

The commonly accepted distribution of actin, myosin, and tropomyosin in striated muscle as deduced from extraction experiments of a more or less selective nature is shown in Fig. 19, reproduced from a review by Perry.[57] However, there are indications that the picture is not quite so simple, especially in relation to the distribution of myosin and to the changes in the density distribution within the sarcomere accompanying shortening. According to the concept of Huxley and Hanson,[9] the actin and myosin are in the form of two separate sets of interdigitating filaments. Shortening is accomplished not by contraction of the filaments themselves, but rather by a mutual sliding action. As the *I* bands disappear, the ends of the myosin filaments pile up against the *Z* band to produce the well-known contraction bands (C_z bands). However, there are many indications in the literature (e.g., Hodge[58]) which suggest that at least some of the material contributing to *A*-band density in the relaxed myofibril is capable of actual migration to the *Z* bands (and possibly to the *M* bands)

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

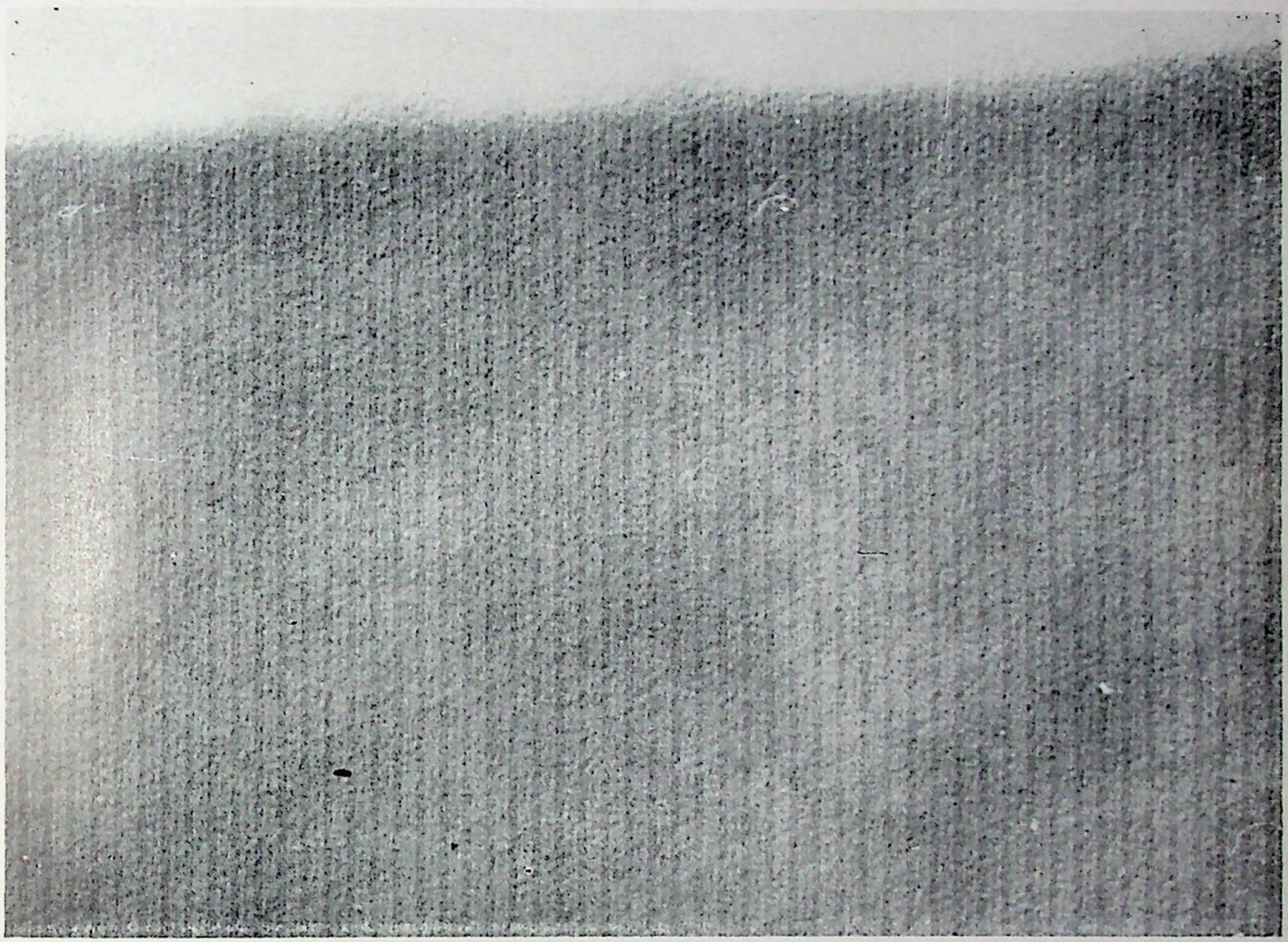


FIG. 15. Electron micrograph of a large flat paramyosin "crystal" obtained by reducing the ionic strength of an approximately neutral paramyosin solution, stained with PTA. The axial period is about 145 A. ×190 000.

independently of any sliding of formed myosin elements. Possibly, such substances act as moderators, activators, or inhibitors in controlling the orderly interaction of the actin, myosin, and other macromolecular components.

There is evidence to indicate that the myosin content of the *A* band is not homogeneous with respect to its extractibility. Thus, on extraction with Guba-Straub-ATP solution, a band of appreciable density remains on both sides of the *M* band[59] (Fig. 20), which can be removed only by further extraction with a pyrophosphate solution.[60] Furthermore, it has been shown that appreciable amounts of components other than myosin are extracted from the *A* band by these procedures.[61] The distribution of birefringence in the sarcomere as found by Inoué and Szent-Györgyi[62] using high-resolution polarization-optical methods coupled with extraction and replating experiments is also at variance with the simple distribution of actin and myosin as postulated by Huxley and Hanson.[63]

The polarization-optical results are more in harmony with the very interesting results of Holtzer and Marshall[64] using the technique of "fluorescent antibody analysis" to determine the localization of the various muscle components. The distributions of actin and myosin found with this technique are essentially in accord with that shown in Fig. 19. However, when using antibodies specific for LMM and HMM, the very curious result emerges that these two components appear to be localized independently of each other within the $A+H$ region of the sarcomere. The LMM appears to be located preferentially in the *A* bands proper, the HMM in a complementary pattern consisting of two bands, one on each side of and adjacent to the *M* band. More recently, Holtzer and Szent-Györgyi[65] have confirmed these observations and demonstrated that there is no physical hindrance to penetration of the antibody into the myofibril, by carrying out experiments involving extraction with KI solutions. In the control, this treatment removes most of the protein of the myofibril. However, pretreatment of a myofibril with a specific antibody prevents the extraction of the particular protein under consideration. The independent but somewhat overlapping distributions found for LMM and HMM by the above methods have also been confirmed by Inoué and Szent-Györgyi,[62] who first selectively extracted for myosin and then exposed the

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

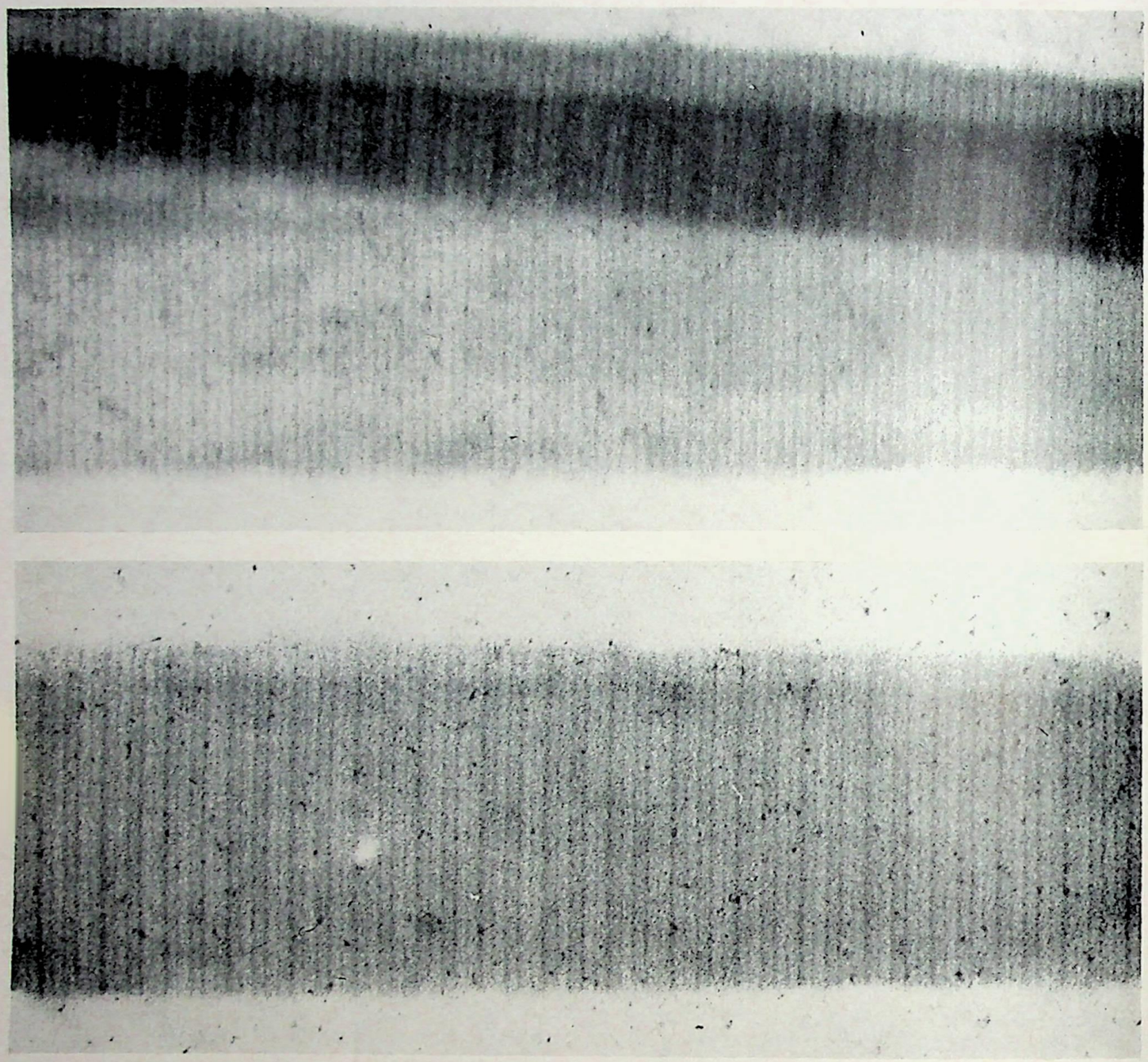


FIG. 16. Two fibrils from the same preparation as Fig. 15. The upper one (a) shows a simple period of about 145 A with a superposed apparently isomorphous structure having a repeat of 725 A, the lower one (b) shows a more complex pattern in which the fundamental repeat period is 5×145 A=725 A. Stained with PTA. Similar patterns have been observed by Hanson *et al.*[56] ×125 000.

myofibrils to LMM and HMM solutions before polarization-optical examination. They found birefringent bands appearing in positions consistent with the fluorescent-antibody results. This apparent independence of at least some of the LMM and HMM within the myofibril recalls the observations of metabolic independence mentioned earlier, and raises the question of whether or not some kind of interaction or reaction of LMM with HMM is involved in the process of contraction.

PERSPECTIVES

The sliding filament model of muscle is based, to a considerable extent, on the observation in the electron microscope of two sets of filaments in the *A* bands, identified by Huxley and Hanson as actin and myosin. However, these filaments have been observed, of necessity, only in thin sections of muscle subjected to the procedures involved in fixation, dehydration, and embedding, so that the original state in the living, resting muscle fiber is a matter of inference. The danger of extrapolation from such observations is illustrated by comparing the low-angle x-ray diffraction data with the results of electron microscopy for muscle fixed (a) in the fresh state and (b) after glycerination. Living muscle gives a low-angle equatorial x-ray diffraction pattern consisting of two reflections corresponding to rod-like elements about 450 A apart in a hexagonal array,[67]

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

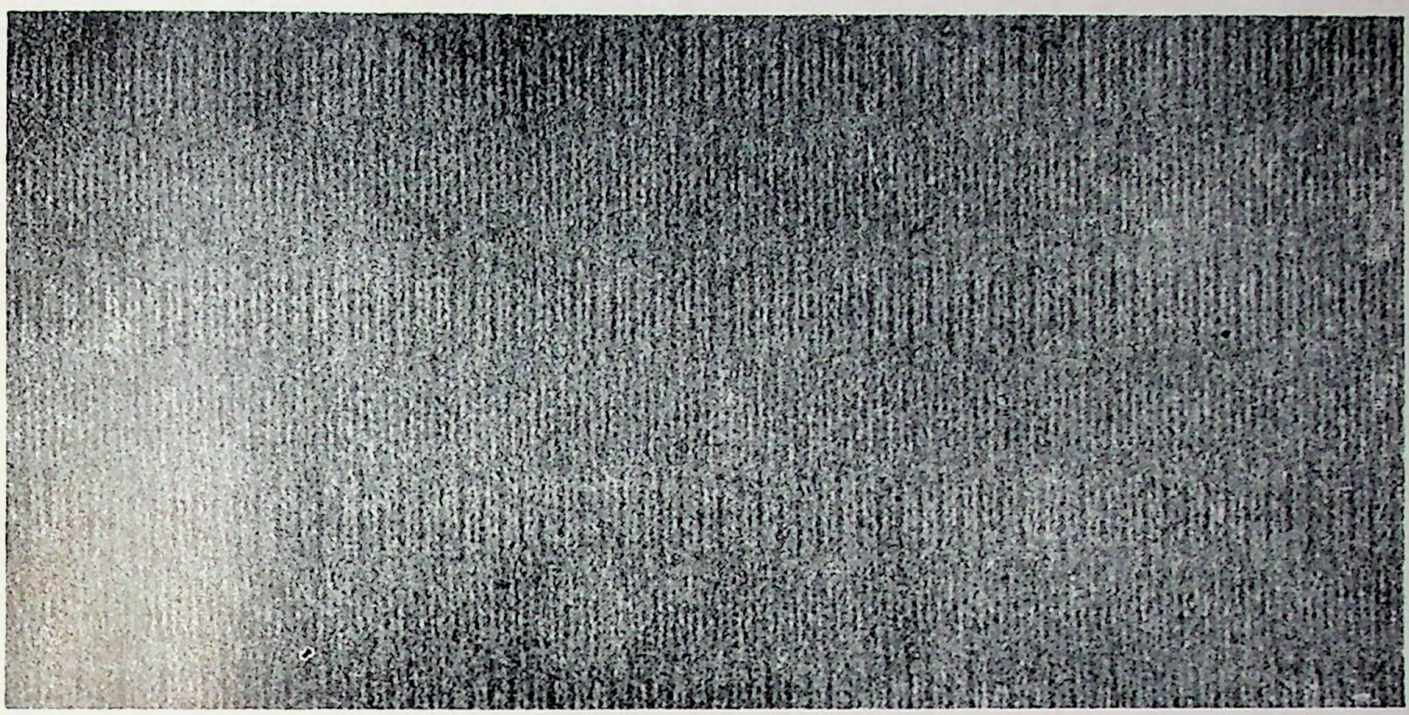


FIG. 17. Paramyosin crystals from the same preparation as Fig. 15, showing a rhythmic transition from a 145-A repeat structure to a *ca* 70-A repeat structure and vice versa. The author is grateful to J. L. Farrant for his suggestion that this pattern could result from the superposition (with small angular displacement of fiber axes) of two paramyosin crystals with 145-A spacings similar to that shown in Fig. 15. ×160 000.

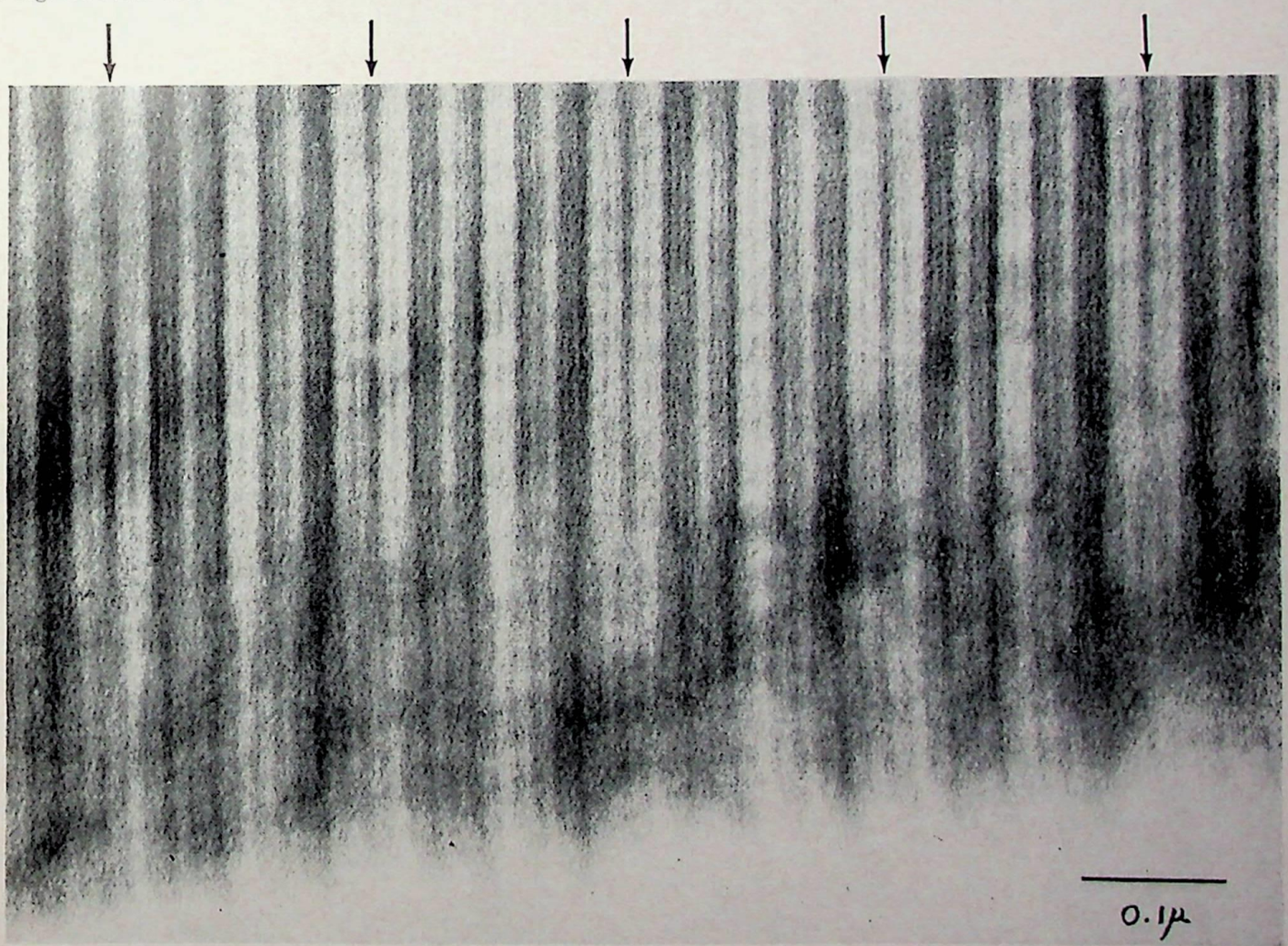


FIG. 18. Paramyosin "crystal" obtained by dilution of a solution of paramyosin in neutral high ionic strength phosphate buffer and stained with PTA, showing a complex band pattern of the type described by Locker and Schmitt[55] with an axial repeat period of about 1800 A (indicated by arrows at top), and a region of transition from the 1800-A repeat structure to one in which the 145-A spacing is accentuated. ×180 000.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

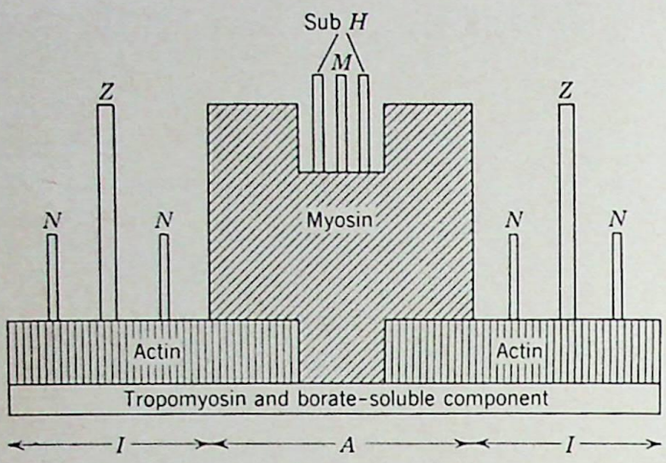


FIG. 19. Diagram illustrating the distribution of the protein components of the myofibril within the sarcomere. The height of the shaded areas represents the protein density at any point along the myofibril axis [from S. V. Perry, Physiol. Rev. **36**, 1 (1956)].

with a strong first order and with a relatively weak second order. If the ATP is removed from the muscle (i.e., in glycerinated muscle or muscle in rigor), a striking reversal of the intensities of these two reflections takes place, the second order becoming very strong.[67] This indicates that, in the absence of ATP, a region of high electron density is present at a position intermediate between the primary rod-like elements, while, when ATP is present, the electron density between these elements is relatively uniform. The electron microscope results are at variance with this, for the two sets of filaments can be observed equally well in muscle fixed in the fresh state or after glycerination.[37] It thus is necessary to conclude that one (or more) of the processes involved in fixation, dehydration, and embedding causes the muscle to go into a rigor-like state, and consequently, that the density distributions observed in electron micrographs do not correspond to those present in the living, resting muscle fiber. The results clearly demonstrate the need for a thorough correlation by x-ray diffraction of the procedures used in the preparation of specimens for the electron-microscopic examination of muscle.

In the sliding-filament model, the two sets of filaments are postulated to remain at constant length on the basis of evidence derived largely from observations of *A*-band lengths during shortening. Thus, this mechanism does not require configurational changes of the protein components, neither at the α-helix nor at the macromolecular level. No direct evidence on this point has been reported. However, Huxley[67] found no change in the axial spacing measured from the low-angle x-ray diffraction pattern on passive stretching of the muscle, a result which has been interpreted as possibly indicating that no change would accompany shortening. On the other hand, some as yet inconclusive evidence from electron-microscopical observations[32,37] suggests that there is a correlation of the axial period both with the

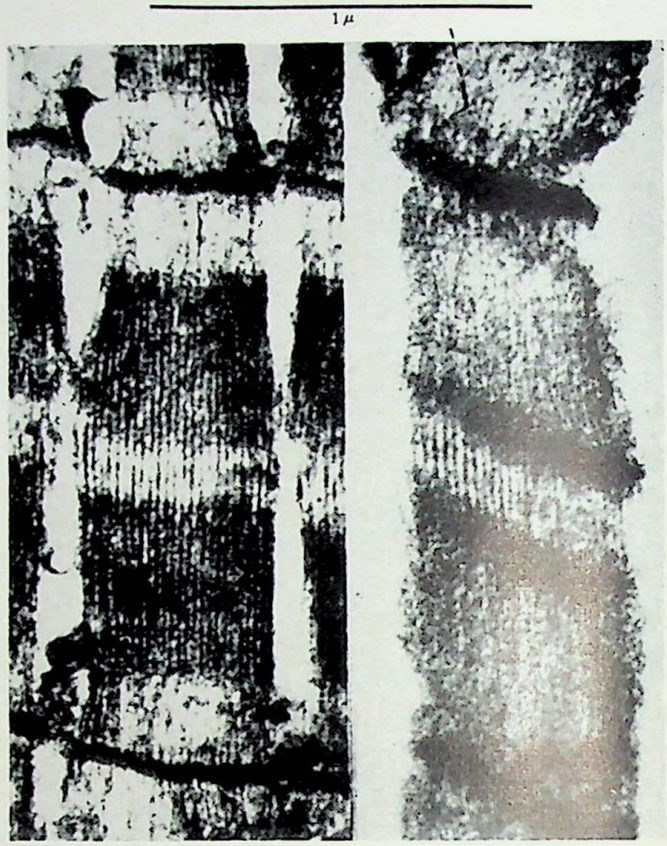


FIG. 20. Electron micrographs of thin sections through (a) intact glycerinated fibril and (b) glycerinated fibril after extraction with Guba-Straub-ATP solution [from J. Hanson and H. E. Huxley, Nature **172**, 530 (1953)].

sarcomere length and with the type of band pattern observed during shortening. Such a variation in the axial period would strongly imply the occurrence of configurational changes during shortening. It is clear that further evidence of a more direct nature is urgently needed in order to settle this matter.

As has been seen, the application of analytical and degradative methods has proved to be of great importance in solving many of the problems associated with muscle structure and function. However, it would seem that the time is ripe for an approach to muscle which considers the whole machine as a complex mixed macromolecular "crystal," capable of changes in physical state in order to convert available chemical energy into mechanical work. Any completely adequate theory of contraction must be able to take into account the interactions between the many and varied components of the myofibril and other contractile units.

BIBLIOGRAPHY

[1] W. A. Engelhardt and M. N. Ljubimowa, Nature **144**, 668 (1939).

[2] Albert Szent-Györgyi, Acta Physiol. Scand. **9**, Suppl., 25 (1945).

[3] Albert Szent-Györgyi, *Chemistry of Muscular Contraction* (Academic Press, Inc., New York, 1951), second edition.

[4] B. B. Marsh, Nature **167**, 1065 (1951).

[5] L. Pauling and R. B. Corey, Nature **171**, 59 (1953).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

[6] L. Pauling and R. B. Corey, Proc. Roy. Soc. (London) **B141**, 21 (1953).
[7] F. H. C. Crick, Acta Cryst. **6**, 689 (1953).
[8] R. S. Bear, J. Am. Chem. Soc. **67**, 1625 (1945).
[9] H. E. Huxley and J. Hanson, Symp. Soc. Exptl. Biol. **9**, 228 (1955).
[10] F. B. Straub, Studies Inst. Med. Chem. Univ. Szeged **2**, 3 (1942).
[11] C. Cohen and A. G. Szent-Györgyi. Paper read at International Biochemistry Congress, Vienna (1958).
[12] W. F. H. M. Mommaerts, J. Biol. Chem. **198**, 445 (1952).
[13] T. C. Tsao, Biochim. et Biophys. Acta **11**, 227 (1953).
[14] K. Laki and W. R. Carroll, Nature **175**, 389 (1955).
[15] A. Holtzer and S. Lowey, J. Am. Chem. Soc. **78**, 5954 (1956).
[16] W. F. H. M. Mommaerts and B. B. Aldrich, Science **126**, 1294 (1957).
[17] J. Gergely and H. Kohler, Conference on the Chemistry of Muscular Contraction, Tokyo (1957).
[18] A. G. Szent-Györgyi, Arch. Biochem. Biophys. **42**, 305 (1953).
[19] J. Gergely, H. Kohler, W. Ritschard, and L. Varga, Abstract of Meeting of the Biophysical Society, Boston (1958), p. 46.
[20] T. C. Tsao, K. Bailey, and G. S. Adair, Biochem. J. **49**, 27 (1951).
[21] C. M. Kay, Biochim. et Biophys. Acta **27**, 469 (1958).
[22] W. T. Astbury and L. C. Spark, Biochim. et Biophys. Acta **1**, 388 (1947).
[23] W. T. Astbury, Nature **160**, 388 (1947).
[24] C. C. Selby and R. S. Bear, J. Biophys. Biochem. Cytol. **2**, 71 (1956).
[25] J. Gergely, J. Biol. Chem. **200**, 543 (1953).
[26] E. Mihályi, J. Biol. Chem. **201**, 197 (1953).
[27] E. Mihályi and A. G. Szent-Györgyi, J. Biol. Chem. **201**, 211 (1953).
[28] J. Gergely, M. A. Gouvea, and D. Karibian, J. Biol. Chem. **212**, 165 (1955).
[29] W. R. Middlebrook, Abstract of Meeting of Biophysical Society, Boston (1958), p. 46.
[30] D. E. Philpott and A. G. Szent-Györgyi, Biochim. et Biophys. Acta **15**, 165 (1954).
[31] C. E. Hall, M. A. Jakus, and F. O. Schmitt, Biol. Bull. **90**, 32 (1946).
[32] M. H. Draper and A. J. Hodge, Australian J. Exptl. Biol. Med. Soc. **27**, 465 (1949).
[33] M. H. Draper and A. J. Hodge, Nature **163**, 576 (1949).
[34] M. H. Draper and A. J. Hodge, Australian J. Exptl. Biol. Med. Soc. **28**, 549 (1950).
[35] S. Leisegang, *Proceedings of the Third International Conference on Electron Microscopy*, London, July 1954 (Royal Microscopical Society, London, 1956), p. 176.
[36] A. J. Hodge, H. E. Huxley, and D. Spiro, J. Exptl. Med. **99**, 201 (1954).
[37] A. J. Hodge, J. Biophys. Biochem. Cytol. **2**, Suppl., 131 (1956).
[38] S. F. Velick, Biochim. et Biophys. Acta. **20**, 228 (1956).
[39] G. Schapira, G. Broun, J. C. Dreyfus, and J. Kruh, Compt. rend. soc. biol. **150**, 944 (1956).
[40] K. Bailey, Biochem. J. **43**, 271 (1948).
[41] A. J. Hodge, A. G. Szent-Györgyi, and C. Cohen (paper in preparation).
[42] M. A. Jakus, J. Biophys. Biochem. Cytol. **2**, Suppl., 243 (1956).
[43] C. Rouiller and E. Fauré-Fremiet, J. Ultrastructure Research **1**, 1 (1957).
[44] C. E. Hall, M. A. Jakus, and F. O. Schmitt, J. Appl. Phys. **16**, 459 (1945).
[45] F. O. Schmitt, R. S. Bear, C. E. Hall, and M. A. Jakus, Ann. New York Acad. Sci. **47**, 799 (1947).
[46] R. S. Bear and C. C. Selby, J. Biophys. Biochem. Cytol. **2**, 55 (1956).
[47] G. F. Elliott, J. Hanson, and J. Lowy, Nature **180**, 1291 (1957).
[48] A. J. Hodge, "Studies on paramyosin; the *in vitro* reconstitution and transformation of periodic structure," Ph.D. thesis, Massachusetts Institute of Technology (1952).
[49] A. J. Hodge, Proc. Natl. Acad. Sci. U. S. **38**, 850 (1952).
[50] F. O. Schmitt, J. Gross, and J. H. Highberger, Proc. Natl. Acad. Sci. U. S. **39**, 459 (1953).
[51] K. Bailey, Publ. Staz. Zool. Napoli **29**, 96 (1956).
[52] K. Bailey, Biochim. et Biophys. Acta **24**, 612 (1957).
[53] D. R. Kominz, F. Saad, and K. Laki, Nature **179**, 206 (1957).
[54] A. G. Szent-Györgyi (personal communication, 1958).
[55] R. H. Locker and F. O. Schmitt, J. Biophys. Biochem. Cytol. **3**, 889 (1957).
[56] W. Moffitt, J. Chem. Phys. **25**, 467 (1956).
[57] S. V. Perry, Physiol. Rev. **36**, 1 (1956).
[58] A. J. Hodge, J. Biophys. Biochem. Cytol. **1**, 361 (1955).
[59] J. Hanson and H. E. Huxley, Nature **172**, 530 (1953).
[60] H. E. Huxley and J. Hanson, Nature **173**, 973 (1954).
[61] A. G. Szent-Györgyi, D. Mazia, and Albert Szent-Györgyi, Biochim. et Biophys. Acta **16**, 339 (1955).
[62] S. Inoué and A. G. Szent-Györgyi (unpublished data).
[63] H. E. Huxley and J. Hanson, Biochim. et Biophys. Acta **23**, 229 (1957).
[64] H. Holtzer and J. M. Marshall, results presented at the Meeting of the Biophysical Society, Boston (1958).
[65] H. Holtzer and A. G. Szent-Györgyi, (unpublished data).
[66] J. Hanson, J. Lowy, H. E. Huxley, K. Bailey, C. M. Kay, and J. C. Ruegg, Nature **180**, 1134 (1957).
[67] H. E. Huxley, Proc. Roy. Soc. (London) **B141**, 59 (1953).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

46
Mechanisms of Muscle Contraction

MANUEL F. MORALES

Department of Biochemistry, Dartmouth Medical School, Hanover, New Hampshire

THIS article discusses current thought on the fundamental problem of muscular contraction, i.e., on the nature of the molecular transducer which converts chemical free energy into mechanical work. Structural observations (Bennett, p. 394; Inoué, p. 402; Hodge, p. 409) have greatly influenced this thought, but most current hypotheses are based to an even greater extent upon the discovery by Engelhardt and Ljubimova that the essential features of muscle action can be reproduced *in vitro*, with fairly pure materials extracted from muscle. This work provided such a far-reaching simplification that many researchers have been impelled to discard the system as a whole and to concentrate on figuring out how the *in vitro* system works. Fortunately, not everyone feels this way.

The simplest system which shows muscle-like contractility is one wherein a thread spun from a protein extracted from muscle is suspended in a neutral medium of low ionic strength. If one places, in this medium, mM Mg^{++} and mM ATP, the thread contracts and does external work. Reciprocally, the thread catalyzes the hydrolysis of ATP to ADP and P—an exergonic reaction[1–4] which is considered to "drive" many biological processes. The question is: How—in a molecular sense—are the hydrolysis and the shortening "coupled"? To give an impression of current thought on this question, some theories or "models" of the process are summarized and passing mention is made of the observations which inspired them. These particular models were chosen, in part because they are among the more reasonable ones, and in part because they were developed from ideas and techniques discussed by Doty (p. 107), Zimm (p. 123), and Rice (p. 69).

In 1948, Kirkwood[5] suggested that the contractile element in muscle might be a flexible polyelectrolyte whose charge is modulated by phosphorylation of its serine residues by ATP. Actually, his was a scheme for relaxation rather than one for contraction, but his basic idea was all-important. At the time, Botts was making a thermoelastic analysis of threads spun from contractile protein, and she came to the conclusion that both the energy and the entropy decreased with increasing length.[6,7] This is the behavior expected of a polyelectrolyte bearing a net charge. Later, impressed by the high charge which ATP has in neutral solution, by the essentiality of Mg^{++} for contraction, and by the fact that Mg^{++} could bind so tightly to the protein as to shift its isoelectric point (I.P.) beyond *p*H 9, we suggested[6,7] that the contractile element of a thread might be a positively charged polyelectrolyte whose charge could be reduced merely by the adsorption of ATP^{4-} in the *first* step of the ATPase process (Fig. 1). This neutralization would lead then to contraction, because, in shortening, the element would—for various reasons—gain entropy. Such a proposition relegated the *hydrolytic* step of the ATPase process to a minor role, a feature that greatly provoked biochemists; however, a conjoint thermodynamic analysis with Hill[8], and later a statistical-mechanical investigation of our model by Hill[9] convinced us that the model is quite acceptable from the point of view of energetics.

The model in question purported to be a model only of the contractile element in muscle. To make the model fit the facts about the *in situ* organization, other assumptions had to be made. In 1946, Schmitt and his associates[10] provided evidence that, in the sarcomeres of uncontracted muscle, the filamentous elements apparently were already extended. Since it was known[11] that these elements exert no tension in passively stretched muscle, I suggested[12] that, at rest, the contractile elements from either end of the sarcomere must interdigitate, but that cross-bonds between such elements must remain absent until the moment of excitation. The main concern, however, is—and was—to see if the model of the *contractile element* is correct, for it is in that device that transduction occurs. In its favor, we and others have mobilized additional evidence of which the following is representative.

Consider the Michaelis-Menten equation which relates the steady-state rate of substrate degradation to the concentration of substrate. It has been shown[13] that the enzymatic activity of the muscle protein system obeys this equation very well. This makes it possible to extract from experimental data the two Michaelis-Menten parameters. The one of interest here is the reciprocal Michaelis constant;

$$\bar{K}=k_1/(k_{-1}+k_2),$$

in the scheme,

$$E+S \underset{k_{-1}}{\overset{K_1}{\rightleftarrows}} X \overset{k_2}{\rightarrow} E+P.$$

There exist simple ways[14] of ascertaining whether $k_{-1}\gg k_2$; i.e., whether $\bar{K}$ can be considered identical with the equilibrium constant of substrate-to-enzyme binding. When this identity is legitimate,

$$\Delta F^0_{\text{binding}}=-RT\ln\bar{K}.$$

Many factors are known which influence the effectiveness with which the ATP structure can bring about contraction; for example, variations in ionic strength, in $[Mg^{++}]$, $[Ca^{++}]$, susbtitutions in the ring (thus

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

generating other members of the nucleotide family), or elimination of the ring (TPP). As a rule, these variations also influence the ATPase activity, and, therefore, one can obtain also a set of $\bar{K}$ values from chemical measurements alone. When one compares *mechanical* measurements of contractant effectiveness with values of $\bar{K}$, there is a fairly good correlation, but in many instances there is no correlation whatever with maximum rate of ATPase activity (as there should be if the tension generator were "geared" to ATP hydrolysis).

Although there are other indications of polyelectrolyte behavior in contractile protein—for instance, the fact that *anions* totally unrelated to ATP, e.g., I^-, SCN^-, $Fe(CN)^{4-}{}_6$, also bring about contraction[15,16]—one additional experimental foundation of the model can be discussed briefly. To do this, and to prepare for the description of other models, the term "contractile protein" has to be defined. The best preparation, from the point of view of work performance, is obtained by extracting directly from muscle mince, according to a well-known recipe, a system referred to as "myosin-B." A contractile system, however, can be obtained also by complexing two separately extracted proteins. One protein is fibrous, exhibits ATPase activity, and is called simply "myosin." The other is globular, contains bound nucleotide, and is called "actin." The contractile complex of the two is called "actomyosin." For the moment, the question is left open as to whether or not myosin-B is identical with actomyosin. Up to this point, myosin-B has been considered in its precipitated (e.g., thread) form, at low ionic strength. If the ionic strength is raised to a few tenths molar, myosin-B dissolves, only then becoming accessible to the traditional tools of physical biochemistry. Electrophoretic examination shows that in such a state the myosin-B particles are anionic. The dissolved system still responds to ATP, by reducing its viscosity, and by lowering its optical turbidity, etc. It has been thought generally that, under these conditions, the underlying structural change is a dissociation of particles into actin and myosin. On this account, the interpretation which we placed on our early light-scattering experiments[17]—that the myosin-B particles were inflating on ATP adsorption—met with something less than global enthusiasm. New investigations[18,19] were carried out in our laboratory, and also while guests in that of Schachman; this new work was sparked principally by Gellert and von Hippel.

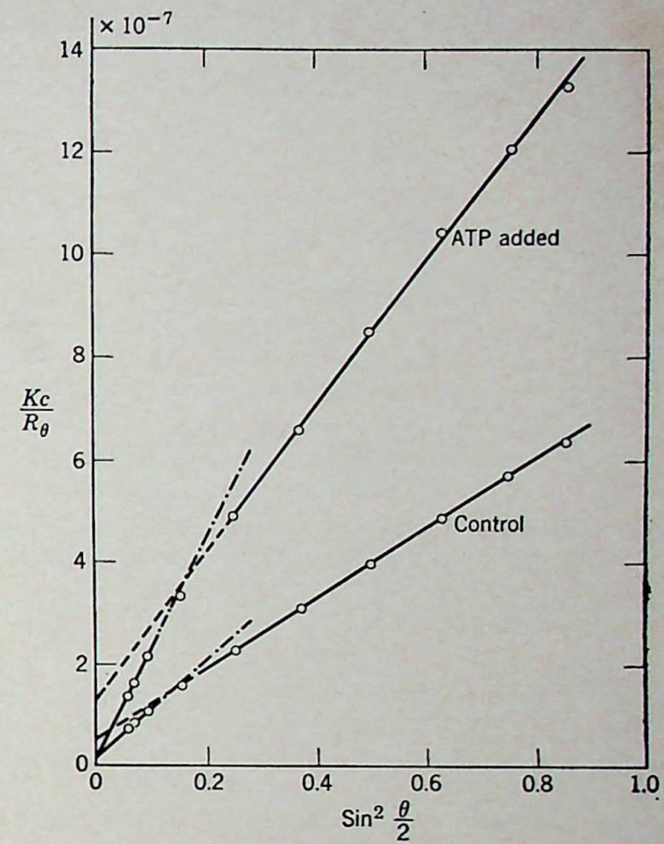


FIG. 2. Zimm plot of the light scattered from a myosin-B solution without and with ATP. Extrapolation from the points obtained at small angles gives an intercept which does not change on ATP addition, thus indicating an essentially constant weight-average molecular weight. ATP addition does increase the slope, however, thus indicating (see text) that at least some particles are "inflating."

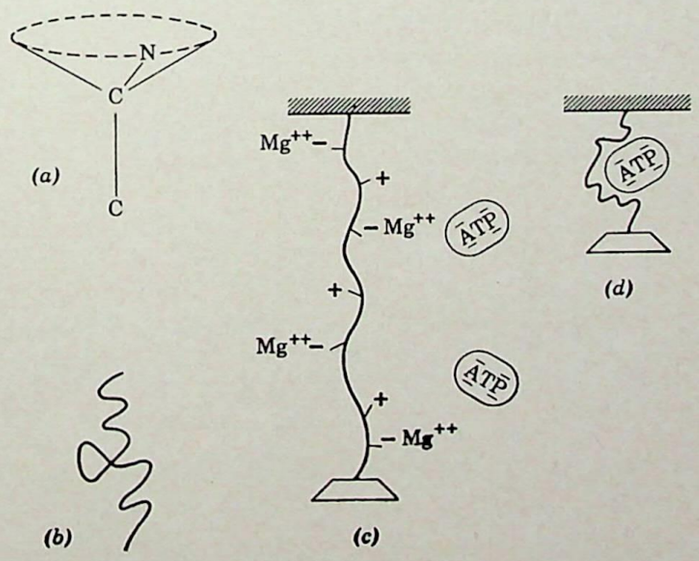


FIG. 1. Elements of a polyelectrolyte model of contraction. It is assumed that the molecular chain has some internal flexibility owing to a free rotation at at least some points (a). In the absence of other effects, the chain then assumes a coiled form which maximizes its configurational entropy (b). However, if the chain bears a net electrostatic charge, it will tend to extend out owing to repulsion between charges. Mg-myosinate might be so extended by a net positive charge in some of its regions (c). If so, adsorption of quadrivalent anions of ATP in the *first* step (complex formation) of the enzymatic hydrolysis would discharge the chain. An effectively neutral chain (d) would shorten toward the configuration of the unperturbed coil (a).

Figure 2, from this work, shows a "Zimm plot" (cf. Doty, p. 61) of data obtained on pristine protein, using the best techniques possible. Only the zero-concentration points are shown because there is ample evidence that the second virial coefficient is negligible. On a plot of this type, the reciprocal intercept gives the weight-average molecular weight ($\bar{M}_w$) and that the initial slope gives a higher order average of the square of a particle dimension. Accepting this data at face

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

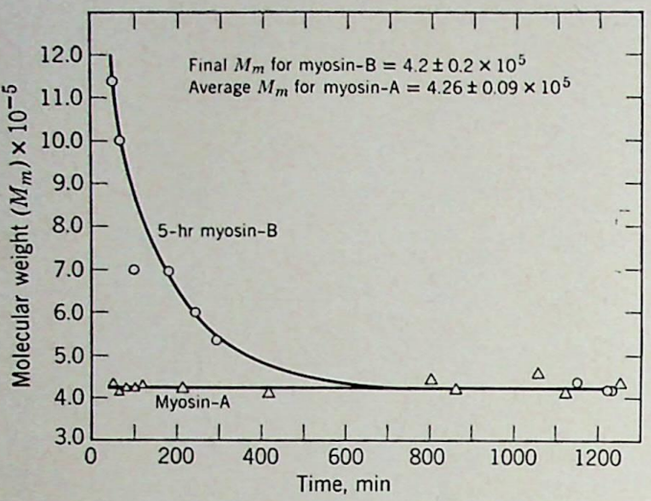


FIG. 3. Molecular weight of myosin (lower curve) and 5-*h* extracted myosin-B (upper curve), measured at the meniscus by the "Archibald" approach to sedimentation equilibrium technique, as a function of time of sedimentation at 4197 rpm and *ca* 5°C. Initial protein concentrations, 0.5 g/100 ml.

value, it is concluded that the addition of a saturating amount of ATP to this system *inflates particles at constant molecular weight*. (The Yang plot, which is more sensitive to changes in $\bar{M}_w$, leads to the same conclusion.) A similar result is obtained using OH^- instead of ATP. Thus, when the protein is in the precipitated form and is cationic, the addition of ATP anions causes contraction; but, when it is anionic, the addition of ATP anions causes inflation. I feel that this constitutes some evidence for the idea that the interaction of ATP with this protein is that of an ion and of a polyelectrolyte. However, conclusions drawn from the light-scattering behavior of a polydisperse system can be very tenuous. It could be, for example, that the solution contains large ATP-inert particles which dominate the scattering at small angles, and smaller particles which simply dissociate (say, into myosin and actin), but whose influence is felt only at larger angles. This actually is suggested in our extrapolations from high-angle measurements. If the amounts and dimensions of these two hypothetical classes of particles were just right, the surviving heavy fraction might be so enriched with *inert* particles that its *average* dimension might appear to increase, even if the particles themselves are not inflating. To block this possibility some excursions in the ultracentrifuge were necessary.

With the ultracentrifuge, we soon found that the myosin-B system was readily separable into light and heavy components. This made it possible to apply the molecular-weight technique developed by Kegeles and Schachman on the basis of the Archibald equations. In Fig. 3 the $\bar{M}_w$, from meniscus measurements, is plotted as a function of sedimentation time. It decreases from very high values (such as are given by light scattering) to an asymptotic value of 4.2×10^5 g. On the other hand, if the same technique is applied to purified *myosin*, from the start one obtains a constant value of 4.2×10^5 g. This result immediately identifies the "light component" in myosin-B as *myosin* and reveals that the molecular weight of myosin is 4.2×10^5 g (until now, this quantity has been very controversial). To explore the heavier and more polydisperse components which generate no schlieren boundary, we turned to the ultraviolet absorbance method. This method gives $C(x)$ rather than $C'(x)$. Figure 4 shows two sample graphs constructed from densitometer measurements on a cell containing myosin-B saturated with ATP; both are plots of absorbance against distance in the direction of centrifugation—the top one corresponding to some time during the acceleration phase and the bottom one to many minutes at speed. By the time the lower measurements had been taken, the heavier components had plummeted to the bottom of the cell, and myosin (the "light component") was beginning to be swept out. From curves such as these, the concentration of myosin was measured. Extrapolating backward in time, allowing for sectorial dilution, the myosin concentration was subtracted from the total concentration to get the concentration of the "heavy components" at the earlier time; for instance, the time at which the upper plot was made. By doing this at various times, the "clearance curve" of the heavy components was inferred, and from it the weight distribution of the various components also, both in the presence and in the absence of saturating concentrations of ATP. The conclusion from these manoeuvers was that the (5-hr extracted) myosin-B system contains about 65% myosin and 35% heavier components, the latter distributed roughly into two classes. On addition of ATP (or PP), about one-third of the total heavy components depolymerizes (yielding

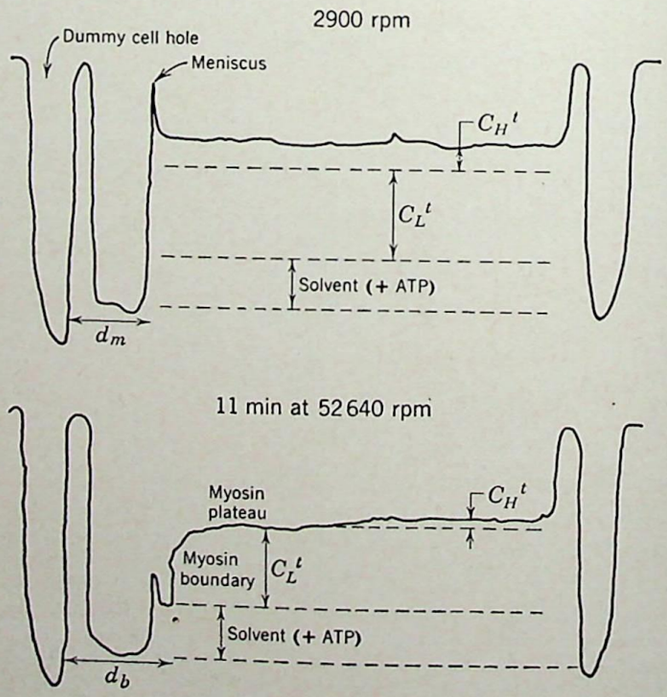


FIG. 4. Absorption of ultraviolet light by an ultracentrifuge cell filled with a solution of myosin-B and ATP, observed at different stages of sedimentation.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

certainly myosin, and possibly actin), while two-thirds seem to remain intact. Based upon these determinations, the light-scattering data were re-examined to determine if the inflation suggested by those measurements might have been apparent only. It turns out very clearly that the inflation must be real, so I feel that one is entitled to say that the nature of the interaction between ATP and the myosin-B system in solution is understandable as an ion-polyelectrolyte effect.

The polyelectrolyte model is a special case of a *class* of models in which it is assumed that the adsorption of ATP—and not its enzymatic cleavage—causes a change in the elasticity of a mechanically continuous protein element. In the same class, for example, are the proposals of Pryor,[20] Flory,[21,22] and Laki[23] in which it is assumed that on adsorption of ATP a "crystallized," perhaps helical, configuration becomes unstable and passes into a random coil. Also in the class is the proposal of Astbury,[24] which envisions passage from the α-configuration into the "super-contracted" configuration.

Another model which also could be constructed from materials found in muscle, but which would operate on a radically different principle has been described by H. Huxley and Hanson and by A. Huxley and Niedergerke[25] and was cast very recently into chemical language by Weber.[26] This model frankly is designed to agree, on the one hand, with certain observations on the organization of proteins within the fibril, and, on the other hand, with the widespread belief that the "coupling" of ATP hydrolysis for the purpose of "driving" another reaction always involves phosphorylated intermediates. The central feature of the Huxleys-Hanson theory[25] of the intrafibrillar organization of proteins is that it adopts, first of all, the "actomyosin" interpretation of the myosin-B system, and considers that over-all shortening is achieved by the motion of actin filaments *relative* to myosin filaments, without invoking the intrinsic contraction of either type of filament. How this sliding motion could result from the interaction of the filaments with ATP is shown in Fig. 5.

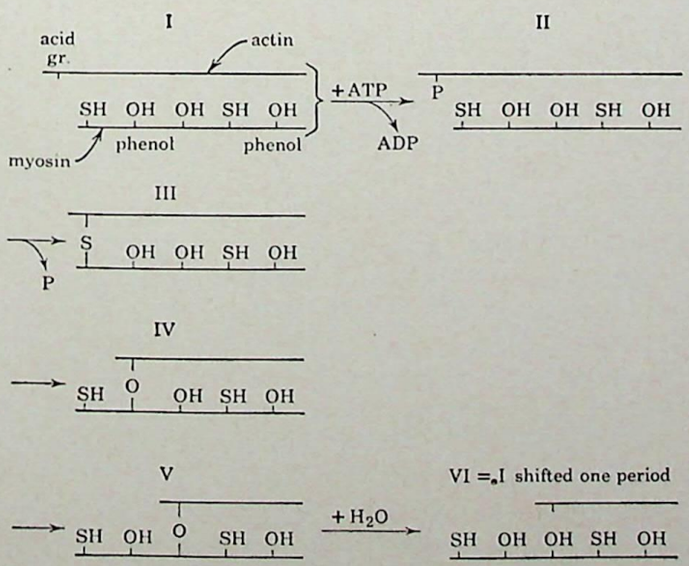


FIG. 5. Weber's[26] proposed chemical scheme to account for the sliding motion assumed in the Huxley-Hanson theory of contraction.

In schemes of this type it is necessary to explain how it is that the filaments slide in one direction only and not in the other, since obviously the reaction scheme is quite symmetrical. Weber does not allude to this point, but it has been well-recognized by Huxley,[27] who *formally* invoked a spatially asymmetric probability of reaction without pretending to give it a chemical basis.

Key support for the Weber scheme is considered to reside in certain tracer experiments of his colleagues, the Ulbrechts.[28] They reported that myosin-B—but not reconstituted actomyosin, which is also contractile—under conditions in which contraction occurs, catalyzed the back incorporation of labeled phosphorus from ADP into ATP. This suggested existence of the reaction,

$$\text{ATP}+\text{actomyosin} \rightleftarrows \text{actomyosin}\cdot\text{P}+\text{ADP}.$$

Since the Ulbrechts found, however, that fibrils from which myosin had been extracted catalyzed the exchange just as well, Weber reasoned that it must be the *actin* partner which is phosphorylated first; that is,

$$\text{ATP}+\text{actin} \rightleftarrows \text{actin}\cdot\text{P}+\text{ADP}.$$

To get the desired translation, he then assumed that P is extruded with the formation of an actin-myosin bond, which bond then migrates across the enzymatic site. At this point, water attacks the bond and the process starts anew with another ATP. This interesting scheme deserves close attention, for, if correct, it should harmonize much chemical and structural data. At the same time, some of its weaknesses must be considered.

Confirmation of the basic Ulbrecht experiment is not yet forthcoming; indeed, in some quarters, there is concern lest the observed exchange might be the result of contamination by granular ATPases. The plausibility of *actin* phosphorylation by ATP is also in serious doubt. Strohman[29] recently demonstrated phosphorylation of actin by the creatine kinase system, but attempts to phosphorylate actin with ATP itself have shown only exchange of whole nucleotide molecules, and no real phosphorylation. Likewise, Levy and Koshland[30] failed to find O^{18} incorporation into ATP when ATP was incubated with actin in the presence of labeled water. Evidence for a transphosphorylation in just the *reverse* direction (i.e., myosin to actin) actually is suggested in the Levy-Koshland work. These authors found that the orthophosphate produced by myosin ATPase was enriched twofold over what it should have been, had the incorporation of O^{18} occurred only at a single hydrolytic step. Still another purely biochemical difficulty arises from considering the speed required by the proposed hydrolytic reaction in order to account for the observed rate of shortening. From the model by Weber, one would expect an individual enzymatic site to be, say, 5 A or 5×10^{-4} μ long. For esterases which are actually

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

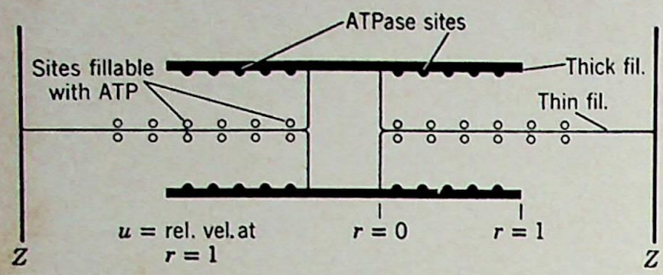


FIG. 6. Scheme for the Podolsky model of contraction. The mechanism for phosphorylating the thin filament (e.g., the creatine kinase-creatine phosphate system) is not included in the diagram.

much faster than myosin-B ATPase, (V_m/E_0) at 25°C is about 10^2 sec^{-1}. Therefore, the relative velocity of filaments moved by the esterase activity should be about 5×10^{-2} μsec^{-1}. Since the myosin filaments are supposed to be closing on the actin filaments from either side, the predicted shortening velocity of the sarcomere is 10^{-1} μsec^{-1}. For an unloaded 3-cm frog muscle contracting at 0°C, Hill[31] reported a shortening speed of 5 cm sec^{-1}. In such a muscle, the sarcomeres, which are 2 μ long, would shorten at a speed of $5/3\times2$ μsec^{-1}, or 3.3 μsec^{-1}, at 25°C this speed easily could reach 10 μsec^{-1}, or 100 times the speed expected from a fast totally unencumbered esterase. To these difficulties should be added two others which stem from the observed length—tension behavior of muscle and myosin-B threads. Both living muscle (single fibers[8]) and myosin-B threads are able to contract until they are 20 to 30% of their original length. But, assuming the forces which bring about the interfilament motion to be short-ranged, it is hard to see how the system ever could contract to less than 50% of its original length. At the other extreme, a living single muscle fiber which is passively stretched and then excited to contraction continues to generate tension until it has been stretched to about twice its rest length. Already at 1.4 times rest length the overlap between the presumably interacting filaments has ceased!

Although I have shown a measure of dissatisfaction with the sliding filament model, I wish at the same time to avoid an impression of blissful contentment with the coiling class of models typified by our polyelectrolyte scheme. To mention but two serious difficulties with coiling models, consider that if myosin is invoked as the polyelectrolyte, one must face up to the difficulty that present measurements indicate that one mole of ATP may influence maximally as much as 10^5 g of myosin. This seems too large a sphere of influence, even for Coulombic forces, and it is the difficulty which has led Szent-Györgyi[32] to invoke very unorthodox means of energy excitation and transmission. Another important shortcoming of the coil model (a shortcoming recently removed for the sliding model[27]) is that, until now, it has never been shown to be consistent with certain heat-work generalizations established by Fenn and Hill for living muscle. In this situation, the proposal recently put forth by Podolsky[33] is most intriguing. His model is basically a variant of the polyelectrolyte model, but its modifications allow it to agree, on the one hand, with some structural observations, and, on the other, with the Fenn-Hill relationships.

Podolsky commences with two empirical equations evolved by Hill,[31] from measurements by Fenn and Hill. These equations describe the heat and work exchanges of a muscle contracting in the neighborhood of its rest length, bearing a load t and shortening at the constant velocity v. The first equation is the "force-velocity" relation (Fig. 6):

$$(\tau+a)(v+b)=5ab;\ a,b,\text{ constants.} \tag{1}$$

The second equation asserts that $-\dot{Q}$, the rate at which the muscle produces heat, is a linear function of v; i.e.,

$$-\dot{Q}=ab+av. \tag{2}$$

From the first law of thermodynamics, it is known that for a system at constant pressure which performs work w in addition to work of expansion,

$$\Delta H=Q-w. \tag{3}$$

If this be a chemical system proceeding under conditions in which ΔH can be expressed as the extent of reaction, ν, multiplied by a (constant) "heat per equivalent of reaction," h, then,

$$h\dot{\nu}=\dot{Q}-\dot{w}. \tag{4}$$

At this point, one has two choices, since Eq. (1) can be solved for either $\tau(v)$ or $v(\tau)$. If one takes $\tau(v)$, combines it with Eq. (2), and substitutes into Eq. (4), one obtains,

$$\dot{\nu}=\dot{\nu}(v,\text{ constants}); \tag{5}$$

if one takes $v(\tau)$, combines it with Eq. (2), and substitutes into Eq. (4) one obtains,

$$\dot{\nu}=\dot{\nu}(\tau,\text{ constants}). \tag{6}$$

Thus, Hill's equations seem to force one to assume that the reaction rate of the net driving reaction depends either upon the shortening velocity or upon the load (analytically, of course, these are equivalent conclusions.) To Podolsky, it has seemed easier to interpret Eq. (5) physically, since a dependence of reaction rate on shortening rate can arise if the two participants in the chemical reaction are attached to structures that are in motion relative one to another. Specifically, he has elaborated the model shown in Fig. 6. At rest, the "thin filament" is supposed to be a positively charged polyelectrolyte at equilibrium length. On excitation, there goes into operation a system capable of depositing on this filament the negative ions of ATP. The thin filament thus develops tension and begins shortening, drawing inwards the two Z membranes. At the same time, the ATPase sites on the thick filaments are free to pick the bound ATP molecules off the thin filament, so that the net number of bound ATP molecules (and,

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

therefore, the tension in the thin filament) will be a balance between the charging rate and the ATPase rate. The ATPase rate now must be considered more closely. The thick filaments do not move. There is, therefore, a *relative* velocity between every point on the thin filament and the point on the thick filament vertically opposite to it. This velocity is zero at the central ends of the thin filaments (where they are supposed to be attached rigidly to the thick filaments), and it rises to a maximum, say u, at the outer ends. If r is a fraction which rises linearly from 0 at the central (fixed) ends of the thin filaments to unity at the outer edges of the thick filaments, it is simplest to assume that, at any position along the thick filament, the relative velocity between opposite points on the two filaments is ru. Since both Z membranes are closing in on the center, each with velocity u,

$$2u/\delta = v \text{ i.e., } u = u(v, \text{ constants}), \tag{7}$$

where δ is the sarcomere rest length and v is the shortening velocity of the whole fiber, expressed in muscle lengths per second. The *chemical* basis for the dependence of reaction rate on shortening rate is introduced in this way (Fig. 7): It is assumed that, as an ATP molecule bound to a thin filament, there can be defined a probability of reaction, P, which depends essentially upon a velocity constant (conventionally defined) and upon the time spent in transit, which in turn depends upon ru. If the distance between bound ATP molecules on the thin filament be l, then an ATPase site, at position corresponding to r, will be presented with passing ATP molecules ru/l times per second; therefore, per second it will carry out $(ru/l)P(ru)$ fruitful reactions, and the entire filament will carry out,

$$\int_0^1 (ru/l)P(ru)\lambda dr \tag{8}$$

fruitful reactions, where λdr is the number of ATPase sites in the element dr. Multiplying expression (8) by the number of filament units per ml of tissue, $\dot{\nu}$ (u, constants) is obtained, and, finally, by means of relation (7),

$$\dot{\nu} = \dot{\nu}(v, \text{ constants}). \tag{9}$$

Thus, on the basis of the Poldolsky model, one can reach the functional dependence (9) which is that called for by the Fenn-Hill relations [Eq. (5)]; moreover, the explicit version of Eq. (9) can be fitted very closely to the experimentally determined version of Eq. (5). The obvious merits of the Podolsky suggestion are three: there is retained what seems to this author the best molecular transducer by far—i.e., the coiling polyelectrolyte (or any of its thermodynamic analogs); there is invoked a structural organization consistent with currently popular ideas; and it is shown that the heat-work behavior of the model would be very like that observed in living muscle.

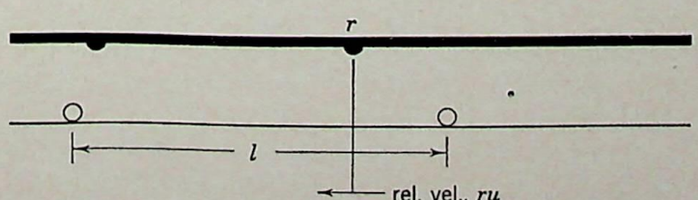


FIG. 7. The motion of an ATP affixed to a thin filament relative to an ATPase site on a thick filament.

To what Podolsky actually has written, I should like to add some opinions of my own. If, indeed, his model is to enjoy consistency with current views on protein deployment, then it must be assumed that his thick filaments are myosin (or its subunits), and that his contractile thin filaments are actin. It is, therefore, perfectly reasonable to assume the existence of ATPase sites on the thick filaments. However, it is not at all obvious that actin could function in the polyelectrolyte sense required by the model in its present form. As already mentioned, actin is a globular protein of monomer weight, about 6×10^4 g. It has an interesting content of bound adenine nucleotide. It is well known that, under certain conditions, the globular (G) form of actin polymerizes into necklace-like aggregates of fibrous (F) actin. As normally prepared, the nucleotide of G-actin is ATP, and of F-actin it is ADP, but the precise mechanistic connection between dephosphorylation and the G-F transformation is unknown. Recently, Oosawa and his collaborators[34,35] carried out an elegant study of the G $\rightleftarrows$ F transformations. They showed them to be cooperative processes, and concluded that the F $\rightarrow$ G transformation is assisted by (a) hydrolysis of bound nucleotide—i.e., $H_2O+G(ATP) \rightarrow F(ADP)+P$— (b) addition of divalent cations such as Mg^{++}, and (c) increase of ionic strength. Naturally, these observations suggest that in G-actin the monomers are kept from aggregating by the electrostatic repulsion between their negative charges; in other words, the dispersed form of actin is charged more negatively than the aggregated form. To the extent that these transformations depend upon the state of the bound nucleotide, they could be driven by interaction with enzyme systems. On the basis of the recent work by Strohman,[29] transphosphorylation from the creatine-phosphate, creatine-kinase system would *disperse* the actin, and transphosphorylation from the dispersed actin to myosin ATPase would *condense* the actin. But this is just opposite to what the Podolsky model requires, since in his model the form of actin with the bound ATP would be assumed to be the *condensing* form, while the form of actin from which the myosin ATPase has removed a P would be assumed to be the *dispersing* form. One way out of this difficulty might be to suppose that *in situ* the actin has sufficient adsorbed cations so that the G $\rightleftarrows$ F transformations are inverted from what they are in free solution, e.g., that the G $\rightarrow$ F condensation is prevented by electrostatic repulsions between *positive* charges on

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

the monomers. Again, not enough is known about intracellular-ion distributions to reject this possibility. To complete this description on a favorable note, I will add that the Podolsky model enjoys a substantial advantage over any coiling model in which it is assumed that myosin is the responding substance, since, if actin is assumed to be the responding substance, the volume of material which would have to be influenced by one molecule of ATP is much smaller and correspondingly more plausible.

Nothing I have read or thought about has weakened my intuitive conviction that the tension-generating device in excited muscle will prove to be a mechanically continuous structure, because any alternative device is apt to run afoul of vector analysis; nor am I ready to relinquish my faith in Coulombic interactions as the most rapidly generated and the most long ranged of the forces that the transducer could employ. On the other hand, I feel utterly dismayed over the experiments which have *not* been done in this field. For example, x-ray diffraction, intrinsic birefringence, and optical-rotation techniques have been available for some time, but one has been content to speculate about small-scale configurational change without trying to establish its occurrence or nonoccurrence in the *in vitro* system. Everything points to the transcending importance of both Ca^{++} and Mg^{++} in the fundamental process—indeed, to the likelihood that it is these ions, not ATP, which are moved in excitation; yet, there is very little certain information on where these ions reside in the fiber. One debates nowadays whether or not ATP or creatine phosphate are dephosphorylated in the twitch, but, so far as I know, little effort is being put into devising optical methods to follow dephosphorylation, and, at present, there are not even the means for blocking one of these reactions and not the other. However, ignorance in these matters will be dispelled only by hard work and certainly not by prolonging this gloomy epilogue.

Added note: Since this manuscript was prepared, W. T. Astbury has informed us that his associates have demonstrated the appearance of the cross-β x-ray diffraction pattern when ATP shortens a film of myosin-B, thus strongly suggesting a real configurational change. Also, F. Buchthal has come forth with electron micrographs which show a wide gap (of nonoverlap) between thin and thick filaments, at muscle lengths at which tension can still be readily demonstrated.

BIBLIOGRAPHY

[1] L. Levintow and A. Meister, J. Biol. Chem. **209**, 265 (1954).
[2] M. Morales, N.M.R.I. Memorandum Report 54-2 (1954).
[3] M. Morales in *International Symposium on Enzymes*, O. H. Gaebler, editor (Academic Press, Inc., New York, 1956), p. 325.
[4] R. J. Podolsky and M. Morales, J. Biol. Chem. **218**, 945 (1956).
[5] J. Riseman and J. G. Kirkwood, J. Am. Chem. Soc. **70**, 2820 (1948).
[6] J. Botts and M. Morales, J. Cellular Comp. Physiol. **37**, 27 (1951).
[7] M. Morales and J. Botts, Arch. Biochem. Biophys. **37**, 283 (1952).
[8] T. L. Hill and M. Morales, Arch. Biochem. Biophys. **37**, 425 (1952).
[9] T. L. Hill, Discussions Faraday Soc. **13**, 132 (1953).
[10] C. E. Hall, M. A. Jakus, and F. O. Schmitt, Biol. Bull. **90**, 32 (1946).
[11] R. W. Ramsey and S. F. Street, J. Cellular Comp. Physiol. **15**, 11 (1940).
[12] M. Morales, Biochim. et Biophys. Acta **2**, 618 (1948).
[13] L. Ouellet, K. J. Laidler, and M. Morales, Arch. Biochem. Biophys. **39**, 37 (1952).
[14] M. Morales, J. Am. Chem. Soc. **77**, 4169 (1955).
[15] L. Churney, Federation Proc. **13**, 26 (1954).
[16] K. Laki and W. J. Bowen, Biochim. et Biophys. Acta **16**, 301 (1955).
[17] J. J. Blum and M. Morales, Arch. Biochem. Biophys. **43**, 176 (1953).
[18] M. F. Gellert, P. H. von Hippel, and M. Morales, J. Am. Chem. Soc. (to be published).
[19] P. H. von Hippel, H. K. Schachman, P. Appel, and M. Morales, Biochim. et Biophys. Acta **28**, 504 (1958).
[20] M. G. M. Pryor, Progr. in Biophys. and Biophys. Chem. **1**, 216 (1950).
[21] P. J. Flory, Science **124**, 53 (1956).
[22] P. J. Flory, J. Am. Chem. Soc. **78**, 5222 (1956).
[23] K. Laki, Discussion at Gordon Conference on Muscle (1958).
[24] W. T. Astbury, Proc. Roy. Soc. (London) **B134**, 303 (1947).
[25] H. Huxley and J. Hanson, and A. F. Huxley and R. Niedergerke, Nature **173**, 971, 979 (1954).
[26] H. H. Weber, *The Motility of Muscle and Cells* (Harvard University Press, Cambridge, 1958).
[27] A. F. Huxley, Progr. in Biophys. and Biophys. Chem. **7**, 257 (1957).
[28] G. Ulbrecht, M. Ulbrecht, and H. J. Wustrow, Biochim. et Biophys. Acta **25**, 110 (1957).
[29] R. C. Strohman, Federation Proc. **17**, 157 (1958).
[30] H. M. Levy and D. E. Koshland, J. Am. Chem. Soc. **80**, 3164 (1958).
[31] A. V. Hill, Proc. Roy. Soc. (London) **B126**, 136 (1938).
[32] Albert Szent-Györgyi, *Bioenergetics* (Academic Press, Inc., New York, 1957).
[33] R. J. Podolsky in Symposium on Cardiac Contractility Ann. N.Y. Acad. Sci. (to be published).
[34] S. Asakura, K. Hotta, N. Imai, T. Ooi, and F. Oosawa in *Proceedings of the Tokyo Conference on Muscle Chemistry*, S. Ebashi, editor (Igaku Shoin, Tokyo, 1957).
[35] F. Oosawa, S. Asakura, K. Hotta, N. Imai, and T. Ooi (to be published).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

REVIEWS OF MODERN PHYSICS VOLUME 31, NUMBER 2 APRIL, 1959

47

Quantitative Studies on Mammalian Cells *in Vitro**

T. T. Puck

Department of Biophysics, University of Colorado Medical Center, Denver 20, Colorado

THE major advances of the science of genetics have come about through the use of two fundamental types of operations. The first and oldest consists in mating of selected, multicellular plants or animals to produce large numbers of offspring, among which the distribution of various inherited characters is examined. This procedure has produced that body of knowledge known as classical genetics and which, beginning systematically with the studies of Mendel, has delineated the processes which sustain and modulate biological heredity in all living forms. This experimental approach, which essentially examines genetic mechanisms in the germ cells, has been extremely successful in elucidating hereditary phenomena in organisms as diverse as *Drosophila* and *Zea mays*, but is difficult to apply to the slowly and less extensively reproducing organisms like the mammals. Thus, while some notable advances have been accomplished, it has not been generally possible to obtain sufficiently large numbers of progeny from selected mammalian matings to provide populations adequate for measurement of many important genetic events.

The second major class of genetic operations involves a more recent technique, which consists in study of the independent micro- and ultramicro-organisms—the free-living cells and the viruses. Here the standard unit is a single cell or particle, and the fundamental operation involves examination of the distribution of genetic traits among the progeny of asexual reproduction. The principal power of this method lies in the vast multiplication factor which it affords. Instead of hundreds or thousands, millions or billions of progeny, all arising from a single individual, can be rapidly produced and quickly scanned for a large variety of genetic characters, so that even rare events can be quantitatively examined. In the last twenty years, such genetic studies on microorganisms have produced a whole new field of knowledge. Sexual methods of genetic analysis are not excluded from these operations but can sometimes be arranged in a step preceding the clonal multiplication of the single cells. In addition, however, new methods for nonsexual genetic analysis have been found which promise to be exceedingly productive. Among the fundamental results which have issued from use of single-cell techniques applied to microorganisms during the last two decades are included: (a) the first systematic demonstration of the direct relationship between genes and enzymes; (b) delineation of many specific, metabolic pathways involved in gene-controlled biosyntheses; (c) demonstration of the direct exchange of genetic materials between cells previously considered to multiply only mitotically; (d) introduction of genes into cells by means of temperate viruses or pure deoxyribonucleic acid; (e) the most accurate measurement of gene-mutation rates; (f) demonstration and quantitative analysis of the processes involved in enzyme induction among genetically competent cells in response to specific environmental stimuli; (g) delineation of the characteristics of linear inheritance in bacteria; (h) systematic use of mitotic crossing-over in certain organisms to localize genes on their chromosomes; and (i) the most detailed mapping at the level of molecular dimensions of the linear sequence of genetic determinants, as accomplished in a bacteriophage chromosome.

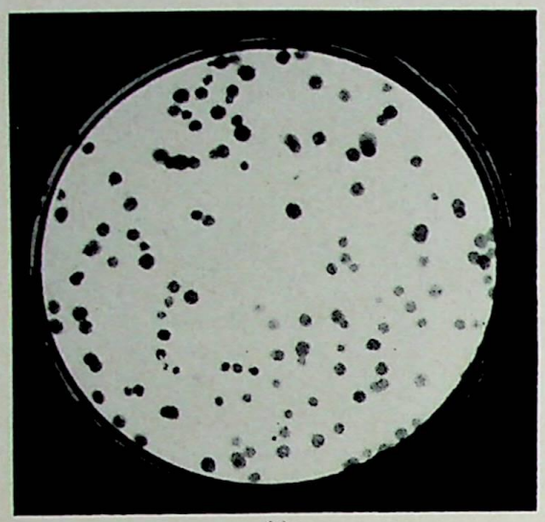
An index of the fruitfulness of this latter approach to the analysis of genetics and related metabolic processes is afforded by the fact that the great majority of the discussion devoted to genetic processes in this biophysical conference has been concerned with developments arising directly from microbial genetics.

In our laboratory, a program was initiated, attempting to develop a methodology that would make possible studies of mammalian cells by means of the techniques of microbial genetics. If successful, such a methodology would provide a new avenue for exploration of mammalian genetic processes among the somatic cells of multicellular organisms, which would complement studies on the genetics of the germ cells achieved by conventional mating studies. In addition, it would afford all of the many kinds of analyses which an unlimited number of progeny provides. Such techniques might facilitate elucidation of the nature of any changes in the genome and phenome, respectively, which are responsible for the characteristic behavior of the various differentiated tissue cells; localization of different genes on the various chromosomes; provision of mutated cells containing specific biochemical blocks for the delineation of various enzymatic steps through quantitative study of specific biosyntheses, like that of hormones and antibodies in specialized cells; search for processes like sexual genetic exchange and transduction in somatic cells, and investigation of the genetic biochemistry of normal and malignant cells.

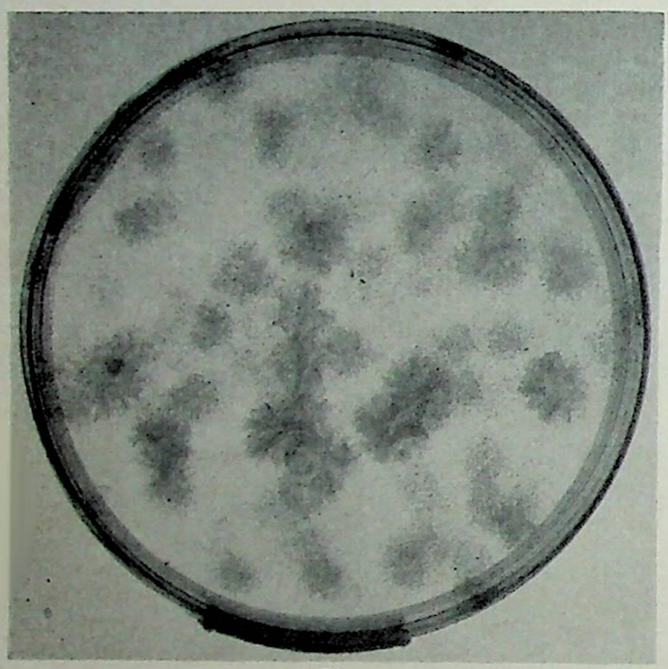
Participating in this program have been a series of devoted co-workers: Steven Cieciura, Harold Fisher, and Philip Marcus carried out studies which formed the basis of their doctoral theses; and D. Morkovin, G. Sato, and N. C. Webb joined this program in a postdoctoral capacity. Chromosome studies have been

* Contribution No. 79 from the Florence R. Sabin Laboratories.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA



(a)



(b)

FIG. 1. (a) Petri dish which was seeded with 100 single cells of the S3 strain of the HeLa culture, an aneuploid which originated in a human cancer. (b) Petri dish seeded with single cells of a euploid human strain. These cells tend more to spread and migrate, and so the colonies tend to run together unless a smaller number of cells is plated or a larger plate is employed.

carried out initially in a collaborative arrangement with Dr. Chu and Dr. Giles at Yale University, and more recently with J. H. Tjio, who joined our laboratory last year. Dr. Arthur Robinson has conducted those aspects of these studies which involved work on human patients.

The first step in this program, which, of course, has drawn a great deal on previously existing tissue-culture methodology, was to develop means for plating single mammalian cells, under conditions such that every single cell develops into a discrete, macroscopic colony.[1] This is essential in order to quantitate growth of single cells of a large population, and to make possible isolation of mutant clones. A simple, quantitative means for growth of single cells into colonies was achieved by a procedure, identical in principle to, and only slightly more complex in practice than, the plating method which constitutes the basis of quantitative microbiology. Figure 1(a) illustrates a representative plating in which 100 cells of the S3 HeLa clone, plated in a Petri dish in nutrient medium, have attached to the glass and produced, within the limits of sampling uncertainty, a discrete macroscopic colony from each cell inoculated. The counting of such colonies is completely objective and highly reliable, so that this methodology permits all of the types of experiments characteristic of quantitative bacteriology to be applied to mammalian cells. Cells taken from organs as diverse as skin, liver, spleen, bone marrow, testis, ovary, kidney, lung, and others, from man and other mammals, and from individuals of a large variety of ages, all respond similarly. It can be concluded that at least large numbers of cells exist in virtually every organ of the mammalian body which retain the potentiality of growth as independent microorganisms, if provided with an adequate physical and chemical environment.

By and large, two general types of cellular and colonial morphologies result from such plating. That illustrated in Fig. 1(a) has been called "epithelial-like" because the cells form compact, tight colonies with relatively smooth, well-defined edges. In our experience so far, the cells that tend to grow stably in this fashion have been polyploid, and usually aneuploid. The other type of colony, which has been called "fibroblast-like" but should perhaps be referred to simply as "stretched," is illustrated in Fig. 1(b), which shows the tendency of the colonies to grow with rough edges, and of the cells to line up as elongated parallel structures. This stretched, needle-like conformation is more characteristic of the euploid cells grown by the methods developed in our laboratory. However, the molecular environment strongly influences the morphology of cells grown in this manner.[1]

The course of developments in microbial genetics has demonstrated the enormous utility of quantitative single-cell plating for the isolation of stocks with rare genetic markers that permit analytical experimentation. Unless each cell can grow in isolation to form a colony, it becomes impossible to apply the highly discriminating screening methods by which millions of cells are subjected to a stressful situation permitting growth of only occasional rare mutants, for otherwise, the exceedingly rare mutant which though potentially is capable of reproduction is not part of a large reproducing population and may not be enabled to express its growth potentiality. An effective single-cell plating technique permits ready recognition of even very rare genetic events, and quantitation of their frequency.

Isolation of mutant clones from such plates is a straightforward process for which a variety of different mechanical methods have been developed. We have established mutant clones which have arisen spontaneously or have been induced as a result of x-irradiation. Among these are included forms with divergent nutri-

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

tional requirements for colony formation, changed colonial morphologies, resistance to destruction by specific viruses like Newcastle disease virus, and differences in chromosomal constitutions (Fig. 2). These mutants have been grown for months or years in continuous culture, during which they have produced astronomical numbers of progeny, and have exhibited stability with respect to their identifying characteristics which is in every way comparable to that of the familiar mutants in bacteria and molds.[2,3]

Attention was next devoted to the development of a defined chemical medium which would promote growth in high efficiency of single mammalian cells, and so make available means for isolating mutants with specific nutritional markers. Many laboratories have contributed studies elucidating small molecular requirements of mammalian cells in tissue cultures utilizing massive cell inocula.[4–6] In studying the growth requirements of individual, isolated cells, we found that single cells of the S3 clonal strain of the HeLa cell form colonies with 100% plating efficiency, when a chemically defined small-molecular medium is supplemented with two macromolecular serum fractions each migrating as a single moving boundary in an electric field. The two proteins of mammalian serum which, under the conditions of these experiments, complete the growth requirements of single cells are serum albumin and an α_1-globulin with a molecular weight of approximately 45 000, which appears to be a glycoprotein in composition. Both substances are necessary for growth of single cells under the specific conditions of our procedures. The functions of these proteins are still under study, but that of the α_1-globulin has been found to be fulfilled completely by the protein fraction named fetuin, an α_1-globulin which constitutes 45% of calf fetal serum protein.[7]

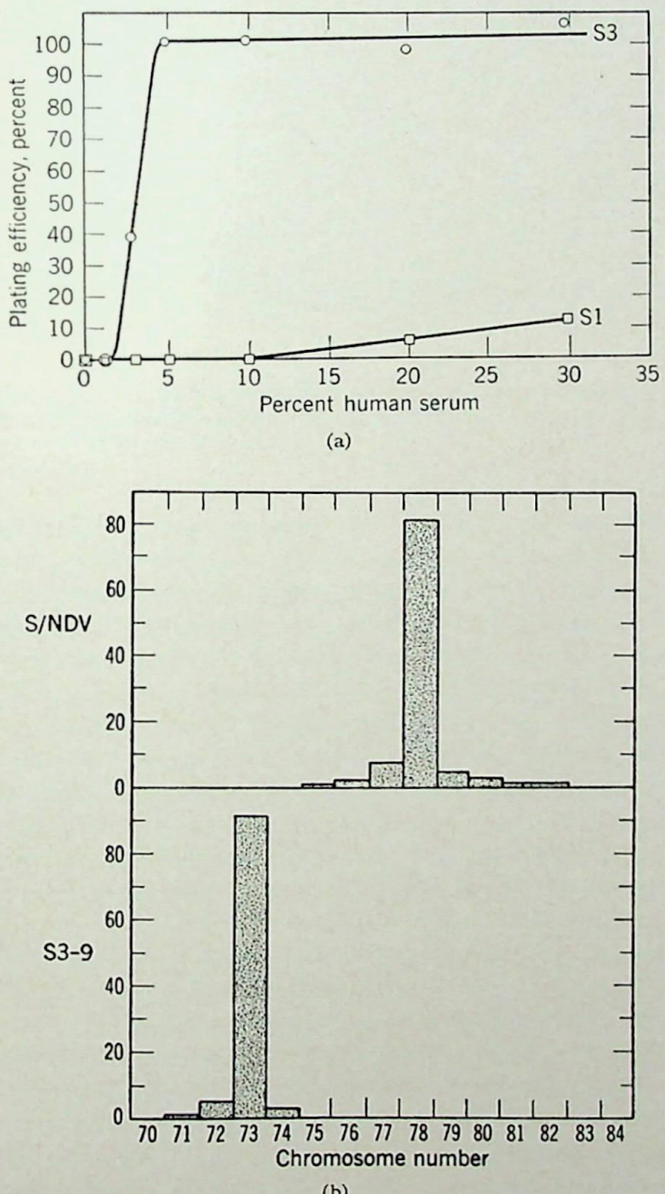
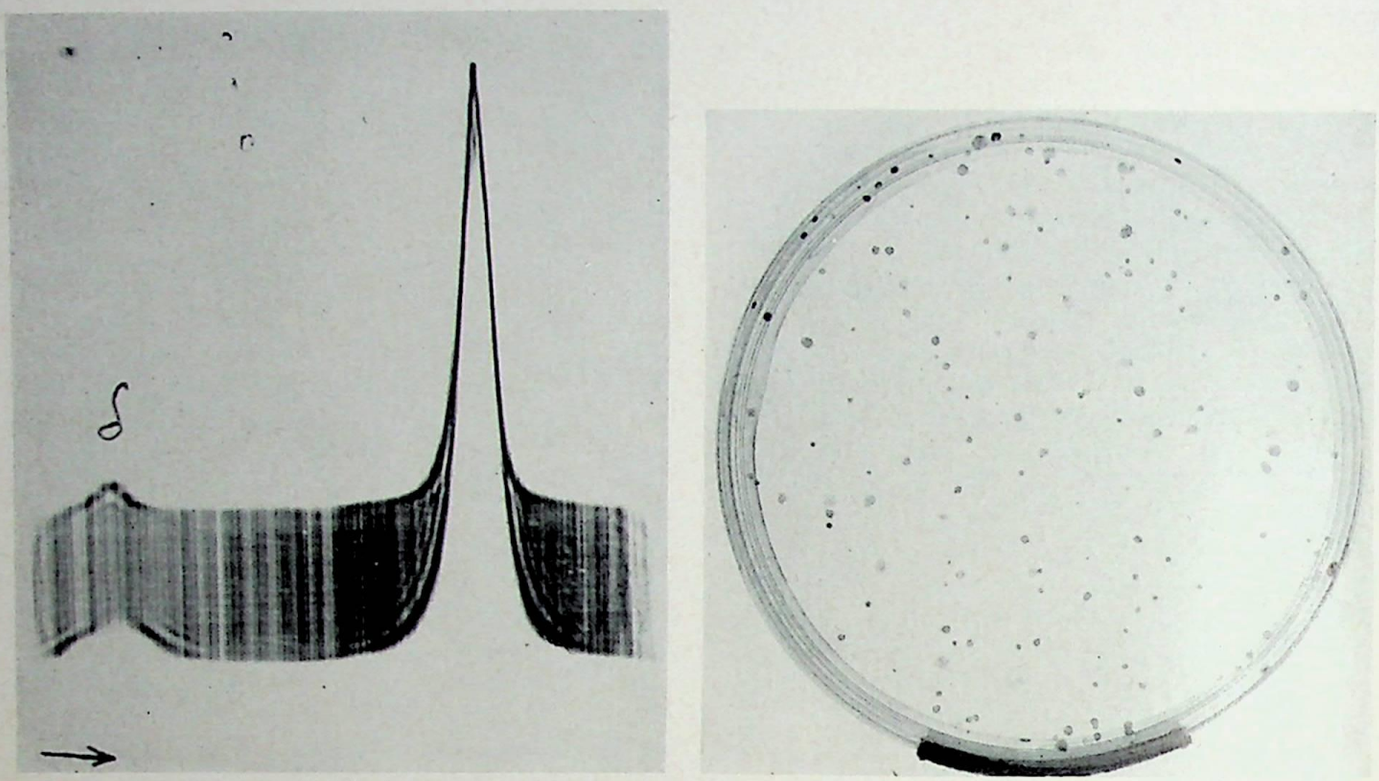


FIG. 2. Demonstration of some representative mutant strains isolated from the HeLa population. (a) Plating efficiency curves in the presence of different amounts of human serum of 2 mutants found to occur spontaneously in the original HeLa population. The cellular and colonial morphologies of these two strains are identical but their growth requirements are different. (b) Distribution of chromosome number in 2 HeLa clones. The clone at the top (*S/NDV*) has the same chromosome number as S3, but is characterized by its resistance to destruction by Newcastle disease virus.

Albumin and fetuin have been purified by repeated precipitation to the point where preparations are approximately 98% homogeneous electrophoretically and ultracentrifugally, although the ultimate purity of such preparations is, of course, difficult to specify with certainty, particularly since fetuin has been demonstrated to be heterogeneous immunologically. A typical electrophoretic pattern is presented in Figs. 3(a), and 3(b) demonstrates the colonial development achieved from inoculation of single cells into a medium containing the purified albumin and fetuin, amino acids, growth factors, salts and glucose. The plating efficiency equals that in serum-containing medium. Experiments on mutants of S3 which exhibit different molecular nutritional requirements are now in progress.

It seems likely that fetuin and albumin may not be required as true metabolites, but rather as conditioning factors which permit establishment of the necessary physicochemical relationships between the cell and surrounding medium. Thus, the fetuin fraction exercises a specific action on the cell membrane, since trypsinized cells added to a medium, complete except for the presence of fetuin, remain as rounded spheres, unattached to the glass surface. With the addition of fetuin to such a medium, the cells rapidly attach to the glass and stretch out in a highly characteristic configuration [Figs. 4(a) and 4(b)]. As little as 1 mg/cc of fetuin can be detected by means of this action. Fetuin has been known to be a powerful antitryptic agent, and at least part of the function of fetuin seems to be antiproteolytic in nature, because this substance prevents the tryptic loosening and rounding of glass-attached cells. The participation in the attachment reaction of a protein not included in the fetuin fraction has been claimed by some workers.[8]

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA



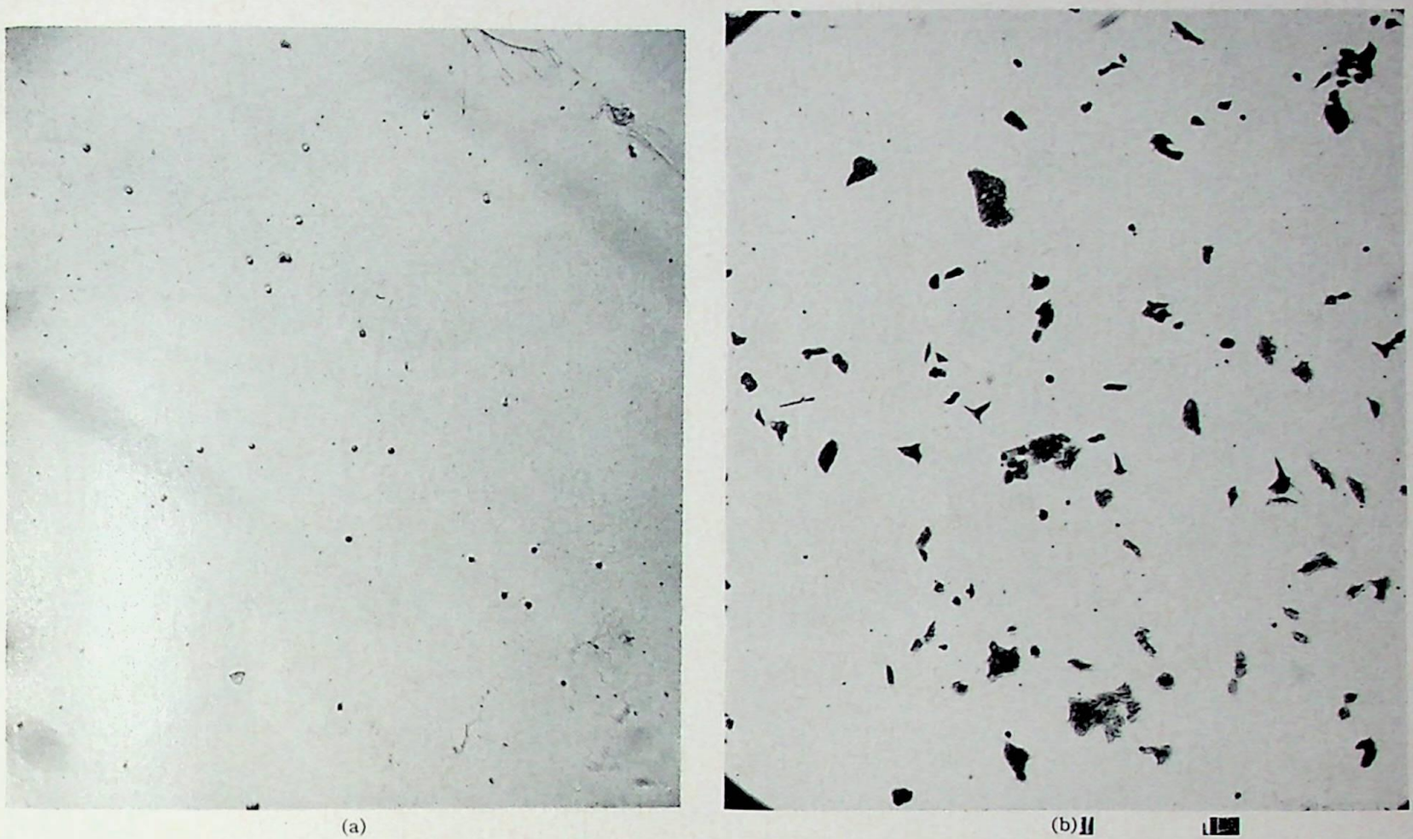
(a) ENZ 18 B1 (b)

FIG. 3. (a) Electrophoretic pattern of fetuin purified by repeated, fractional $(NH_4)_2SO_4$ precipitation from fetal calf serum. (b) Colonies developing on plate seeded with 100 S3 cells, in a medium containing known small molecular components, and purified fetuin and human serum albumin.

In further development of tools for quantitative study of the growth and genetics of the mammalian cell *in vitro*, it became necessary to attempt to control the lability of karyotype which had been demonstrated to affect cells cultivated by standard tissue-culture techniques. Recent studies have shown that virtually every mammalian cell which has been established as a stable tissue-culture strain has a chromosomal constitution radically different from that of the parental species from which it originated.[9,10] Euploid cells were generally found either to stop growth after a very short period in culture, or else to proliferate, but with drastic changes in chromosome number and structure. In order to achieve conditions eliminating this proclivity for aneuploid chromosomal constitution in cell cultures (which renders genetic studies of questionable significance), methodologies were required for simple and reliable monitoring of chromosomal constitution of cells grown *in vitro*. Such methods were developed by Axelrad and McCulloch,[11] by Rothfels and Siminovitch,[12] and by Tjio and the author in our laboratories.[13] When the karyotypes of such cultures could be checked with ease and accuracy, it was possible to find growth conditions which permit karyotype integrity to be maintained. These conditions were secured by development of a screening test for media and regulation of other conditions so as to remove conditions leading to mitotic inhibition, which might induce polyploidy. By careful control of the processes involved in cell dispersion, by regulation of the *p*H and temperature during cell incubation, and by use of pretested media which incorporate all the needed nutritional requirements for cell growth and exclude toxic substances, it became possible to grow cells from normal tissues of man and other mammals, for periods now approaching a year, and for numbers of progeny which have exceeded 10^{25} without change in chromosomal constitution.[14] Figure 5 shows the chromosome complement of typical human male and female cells. These methodologies have permitted characterization of each of the human chromosomes, and particularly of the *X* and *Y* sex-determining chromosomes.[15]

In order to employ human genetic markers with biochemical, immunologic, and pathologic characteristics, a method was devised by which cells from any individual could be simply and reliably introduced into stable growth *in vitro*. A small piece of skin, approximately 10 to 50 mg in mass, is excised from the ventral part of the forearm, dispersed by trypsinization, and grown in the medium which contains fetal calf serum, which was shown to maintain stable karyotype. This method has proved itself reliable in securing actively growing cultures from any individual.[14] It has made possible confirmation of the fact that the human karyotype does, indeed, consist of 46 chromosomes, as first reported by Tjio and Levan.[16] Studies are now under way on cells of individuals with phenyl ketonuria and other genetic diseases, in an attempt to establish

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA



(a) (b)

FIG. 4. (a) Photomicrograph demonstrating the rounded condition of cells, which will shortly be released from their bond to the glass in the absence of any "stretching factor." (b) Photomicrograph demonstrating the stretched condition of the cells in an adequate concentration of purified fetuin.

cytogenetic and biochemical differences between these and normal cells in the *in vitro* cultures. These methodologies are also being employed clinically to determine the chromosomal constitution of persons with various degrees of clinical hermaphroditism and to compare the results so obtained with those utilizing the sex chromatin method devised by Barr and his associates.[14] Thus, an anatomically female patient suffering from ovarian dysgenesis (Turner's syndrome) was shown to possess only 45 chromosomes.

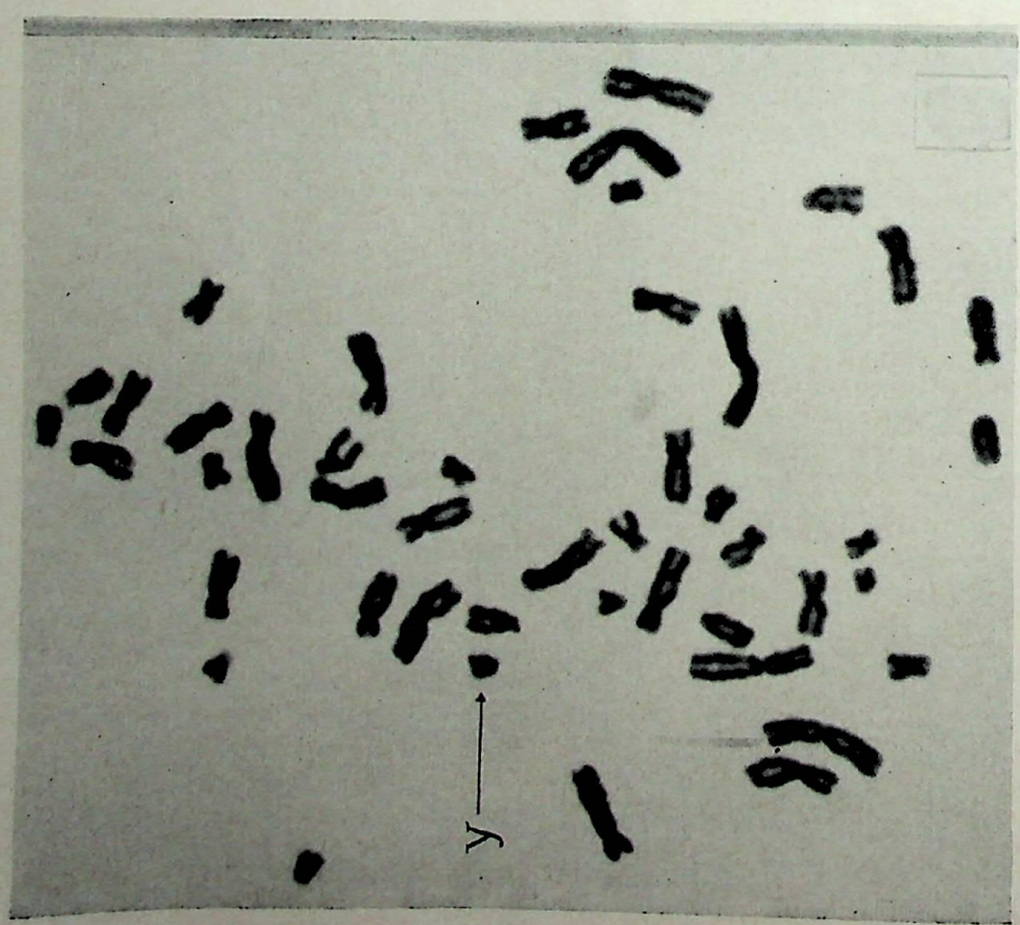
In contrast with the normal human chromosome constitution shown in Fig. 5, Fig. 6 presents that of the aneuploid S3 HeLa cell, a clonal stock which we developed from the culture which Gey and his associates established from a human carcinoma of the cervix.[17] The chromosome number of this clonal strain is 78, which represents a hypotetraploid condition. Studies are in progress comparing the different morphology and behavior of cells which differ in chromosomal number and constitution. Thus, mutant clones of animals like the Chinese hamster have been isolated, which have stemline chromosome numbers of 22 and 23, respectively.

As one example of the kinds of quantitative studies made possible by these techniques, the remainder of this discussion is devoted to analysis of their application to study of the action of high-energy radiation on mammalian cells. Ionizing radiations are capable of disrupting any chemical bond in any molecule of any cell which has absorbed such energy. On the basis of such qualitative considerations alone, one might expect an enormous variety of cellular pathologic actions. Radiobiologic studies on mammalian systems carried out by many laboratories over many years have more than borne out this expectation, demonstrating a bewildering variety of pathological consequences attending various kinds of exposure to radiation of mammals, including phenomena as diverse as gastrointestinal disturbance, drop in the level of blood white cells, epilation, and carcinogenesis. One of the most critical problems in this field involves determination of the degree to which chemical disorganization attending exposure to a given dose of ionizing radiation affects the genome or the phenome, respectively, of the irradiated cell. A quantitative answer to this question would greatly simplify attempts to understand the enormously complex series of events attending irradiation of mammalian cells and of whole animals. In addition, the answer to this question is essential in order to determine the degree to which exposure of mammalian populations to various doses of ionizing radiations will result in random mutations transmissible to succeeding generations.

Physiologic damage to cells in the form of reversible lag periods in cell multiplication, changes in cell permeability and inhibition of many specific enzymatic activities, have been demonstrated in many organisms.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

(b)



(a)

FIG. 5. (a) Chromosome constitution of normal human male cell.[13] (b) Chromosome constitution of normal human female cell.[13]

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

FIG. 6. Chromosome complement of a typical S39 HeLa cell, a subclone of the S3 strain, of human malignant origin, with a stemline number of 78.

In a variety of nonmammalian forms it had also been shown that the cell nucleus is much more sensitive to damage by irradiation than the cytoplasm. Genetic and cytologic examination of radiation effects, particularly in the simpler organisms, has revealed that changes of the genetic structure arising from irradiation include single gene mutations, chromosome breaks, production of areas deficient in chromatin, and occasionally elongated and amorphous chromosomes, these latter presumably arising from cells which had been irradiated during mitosis. At the site of fragmentation, a chromosome exhibits sticky surfaces which can re-attach so that the chromosomes may reconstitute themselves into a configuration more or less closely resembling the normal. In all probability, however, a defect, even though invisible, will persist at the site of breakage unless this happened to lie in a genetically inactive region. All or part of any chromosome which has suffered such a "hit" may, through imperfect restitution, fail to be incorporated in the mitotic apparatus and be lost in subsequent cell divisions. Such a condition can cause genetic imbalance that may destroy the ability of a cell exhibiting it to reproduce indefinitely. In addition, if an unrestituted chromosome divides, the two sister chromatids may unite at their broken ends to form an anaphase bridge that can prevent completion of mitosis. If multiple chromosome hits occur in the same cell, all of the foregoing damaging actions characteristic of single breaks are intensified. Moreover, abnormal restitution of the sticky, broken ends of different chromosomes may occur, producing a variety of complex translocations, such as dicentric chromosomes which can destroy the cell's reproductive capacity.[18]

Earlier studies of the relative sensitivities of the mammalian genome and phenome, respectively, to damage by high-energy radiations at first seemed to favor the thesis that for mammalian cells, in contrast to cells of other organisms, physiologic rather than genetic effects may be most important in the dose range up to and including the mean lethal dose for the whole organism, which is about 400 to 500 r. This supposition appeared to be supported by several kinds of evidence: studies in which massive cell inocula were irradiated in tissue culture with progressive doses of irradiation revealed that permanent loss of cell proliferation did

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

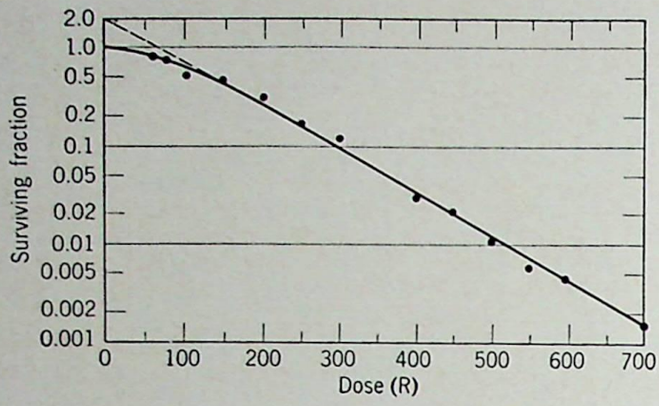


FIG. 7. X-ray survival curve of the ability of single S3 HeLa cells to form colonies as a function of the dose, D. The points can be fitted either by the equation $N/N_0=1-[1-e^{-D/96}]^2$ or by $N/N_0=e^{-D/77}(1+D/77)$, the latter relationship being preferable, since it is based on the model that reproductive death for this polyploid cell requires two or more effective hits anywhere in the chromosome complement. A more refined and comprehensive equation which expresses the contributions to genetic death of haploid, diploid, or polyploid cells of any species by single-hit and multiple-hit radiation damages is presented elsewhere.

not occur until many thousands of roentgens had been administered. Presumably because of the way in which tissue-culture thinking in the past was dominated by the concept that cell proliferation is a property of a cellular community, experiments like this were interpreted as an index of the dose needed to inactivate the cellular reproductive mechanism. Moreover, irradiation of single cells of bacteria, yeasts and protozoa, which might be considered as models for the mammalian cell, indicated that in all these forms, inactivation of reproduction required thousands of roentgens.[19] Hence, since death of a whole mammal can be effected by an exposure of only 400 to 500 roentgens, it appeared possible that mammalian-cell interaction with irradiation proceeds in a manner fundamentally different from that of other forms. While this thesis has been vigorously opposed, especially by geneticists, among whom H. J. Muller has been particularly active, it is still occasionally affirmed in the current scientific literature that very high doses in the neighborhood of many thousands of roentgens are needed to achieve irreversible damage to mammalian cells *in vitro*[20] and the ability of ionizing radiation to produce mutagenesis in mammals is still being challenged.[21]

With the development of the techniques which have been here described, it became evident that the use of cell proliferation from massive cultures cannot be used as an end point to quantitate irreversible effects of radiation on cell reproductive capacity unless the results are calculated as a survival function based on the number of cells irradiated. Since even a single cell remaining intact can proliferate eventually to produce any degree of outgrowth whatever, the dose at which a large but unspecified population fails to produce recognizable outgrowth is meaningless in terms of the dose needed on the average to inactivate the cellular reproductive apparatus. Moreover, since it is characteristic of irradiated cells to exhibit reversible mitotic lags, and to continue to multiply for one or even several generations, but then to produce a microcolony of sterile progeny, use of the mitotic index of the population as a measure of irreversible effects on cellular reproduction is not reliable. However, the use of the single-cell plating procedure makes possible precise determination of survival curves for mammalian cells by procedures exactly like those which had been used for cells of *E. coli*. In this way it is possible to measure accurately the dose required for inactivation of the reproductive mechanism of the individual cells constituting a population. The first such curve obtained is shown in Fig. 7, and approximates a two-hit relationship, with an initial shoulder, followed by a linear exponential drop, which is maintained for doses up to at least 1800 r. Of major importance is the fact that the mean lethal dose, D_0, which is obtained from the slope of the linear part of the curve, was only 96 r. The determination of this curve is completely objective and has been verified by this time in a number of laboratories.

Several features of the behavior of such irradiated cells pointed to damage to the genome as the primary process responsible for reproductive death.[19] The small value of the mean lethal dose for the S3 HeLa cell made it possible to rule out of consideration single-gene mutation as the dominant radiobiological process. The hypothesis was advanced that cell death is a consequence of chromosomal damage resulting from irradiation, and this proposal appeared to give good quantitative fit with the behavior of the system. This interpretation was supported by the following facts. After x-irradiation of single cells with approximately 6 mean lethal doses, the survivors were allowed to form colonies which were then picked and grown into new cell stocks. Examination showed at least 4 out of 5 such strains to be mutated from the original form, exhibiting morphological or nutritional differences, which persisted after long-term growth. Moreover, each of these four strains was revealed to possess a chromosomal constitution grossly altered from that of the original, unirradiated strain (Fig. 8). The 2-hit nature of the curve for the S3 cell was explained on the basis that reproductive death in such cells requires, in the main, at least two independent hits, and therefore appears to be a result of aberration produced by interaction between two separately damaged chromosomes. Other aneuploid cells of human origin, like S3, and with chromosome numbers far in excess of the normal 46, gave x-ray survival curves similar to that of S3.[22]

The fate of the cells whose ability to multiply indefinitely has been destroyed by irradiation is of interest. While some may multiply for a few generations, all appear to retain the ability to metabolize effectively and to carry out complex biosyntheses of macromolecular structures like that of a specific virus, added to the system after irradiation. Under proper environmental conditions, cells of either aneuploid or euploid

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

constitution may, after irradiation, produce giant forms. Thus, these cells which have been destroyed reproductively by irradiation, have maintained intact a large proportion of their other metabolic activities. At least small distortions of these functions probably also exist, but may require subtle means for their demonstration.

A pattern of events similar in some ways, but different in others, was obtained on irradiation of human cells with normal chromosome complements. All of these cells originated in normal human tissues and possessed the typical, elongated configuration commonly associated with euploid human cells.[14] Determination of chromosome constitution on several of these revealed them to have the normal diploid number. All of them behaved identically, and exhibited a mean lethal dose of about 50 r, corresponding to an even greater radiosensitivity than the HeLa cell. The hit number of these survival curves is less than that of the aneuploid HeLa cell, and lies somewhere between 1 and 2, a greater accuracy not yet being available for these cells because of their great radiation sensitivity. Hence, the conclusion may be drawn that euploid human cells are even more radiosensitive than the malignant, aneuploid HeLa.

As a further test of the hypothesis that cell reproductive death, as a result of x-irradiation, arises directly from chromosomal damage, a series of experiments was carried out in which euploid human cells were irradiated, after which the chromosomes were delineated and examined cytologically for direct evidence of aberrations. A preliminary report by Bender,[24] counting

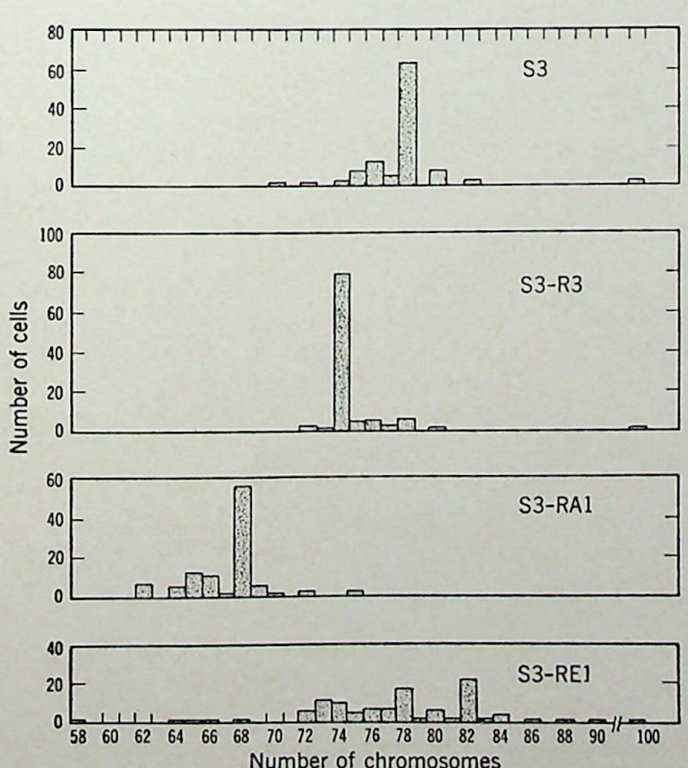


FIG. 8. Chromosome number distribution of the clonal strain S3 (top), and of 3 typical subclones isolated from among the survivors of radiation with 500–700 r (S3R3, S3RA1, and S3RE1). These chromosome analyses were carried out by Chu and Giles[23] in a collaborative study with our laboratory.

TABLE I. Chromosome aberrations at various radiation doses. Single-hit aberrations are chromosomal defects caused by a single ionizing event, and include a complete break in one chromatid only; a break in both chromatids at the same point presumably reflecting a break in the chromosome before it has doubled or a traverse of both chromatids by the same ionizing particle; an achromatic region in which the continuity of the chromosome is uninterrupted, but chromatin has disappeared from a particular area; or the presence of one or more greatly elongated chromosomes trailing "sticky" streamers. Multihit complexes comprise chromosomal aberrations involving interaction between two or more independent hits to 1 or more chromosomes, such as translocated dicentrics and ring chromosomes.

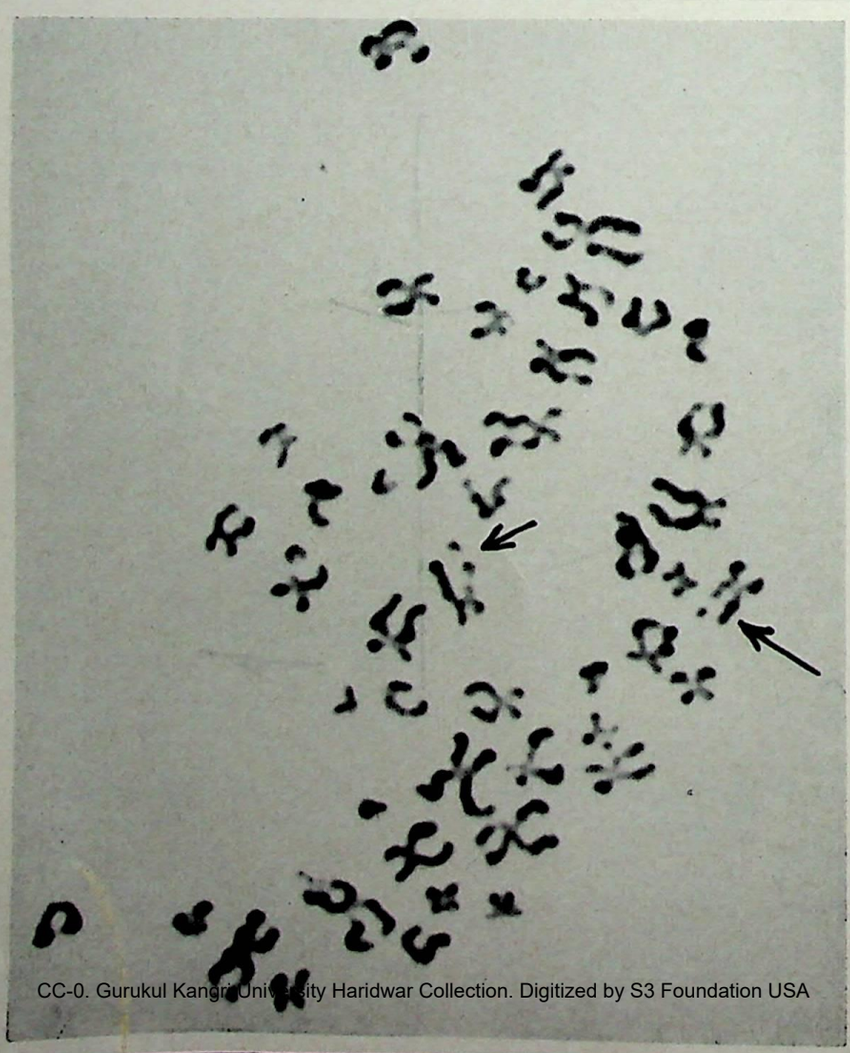
		Single-hit aberrations				
Dose (roentgens)	No. of mitoses scored	Single chromatid break	Double chromatid break	Achromatic regions	Presence of one or more "sticky" chromosomes	Multihit aberrations
0	116	22	1	1	1	1
10	3	0	0	0	0	0
20–25	33	6	1	0	0	0
40–50	20	37	4	0	2	1
75	101	113	23	7	2	14
150	26	26	5	8	4	10
Totals	299	204	34	16	9	26

chromosome breaks induced in euploid human cells grown *in vitro*, suggested that approximately 300 r were required to produce an average of one visible chromosome break per cell. This figure is almost six times higher than the mean lethal dose for these cells. In an attempt to obtain an estimate of the roentgen efficiency of chromosome breaks for euploid human cells, valid under our conditions of study, experiments were carried out in our laboratory in which the conditions of growth of the cells were brought as nearly as possible to optimal values in order to minimize some of the uncertainties owing to mitotic lags. In addition, however, an independent method of arriving at the chromosome-breaking efficiency was employed, based on the fact that the appearance of aberrations due to abnormal chromosome restitutions, can occur only in appreciable numbers at doses beyond the mean lethal dose. Such abnormal restitutions will not disappear, as can single chromosome breaks which can reseal without leaving any visible trace.[25]

Figure 9 presents typical sets of mitotic figures obtained when cells with normal chromosome constitution taken from various tissues of normal human subjects are irradiated *in vitro* at different dose levels. It is obvious that, with cells irradiated with 50 r, single chromosome breaks are readily evident [Fig. 9(b)]. When the dose is increased to 75 r, two effects appear: the number of single breaks is increased; in addition, however, new types of aberrations appear which indicate that, at this dose, multiple chromosome hits in individual cells have become a frequent occurrence. Examples of such complex anomalies formed through interaction of several radiation-damaged chromosomal sites are presented in Figs. 9(c) and 9(d). A summary of this data is presented in Table I. From these figures,

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

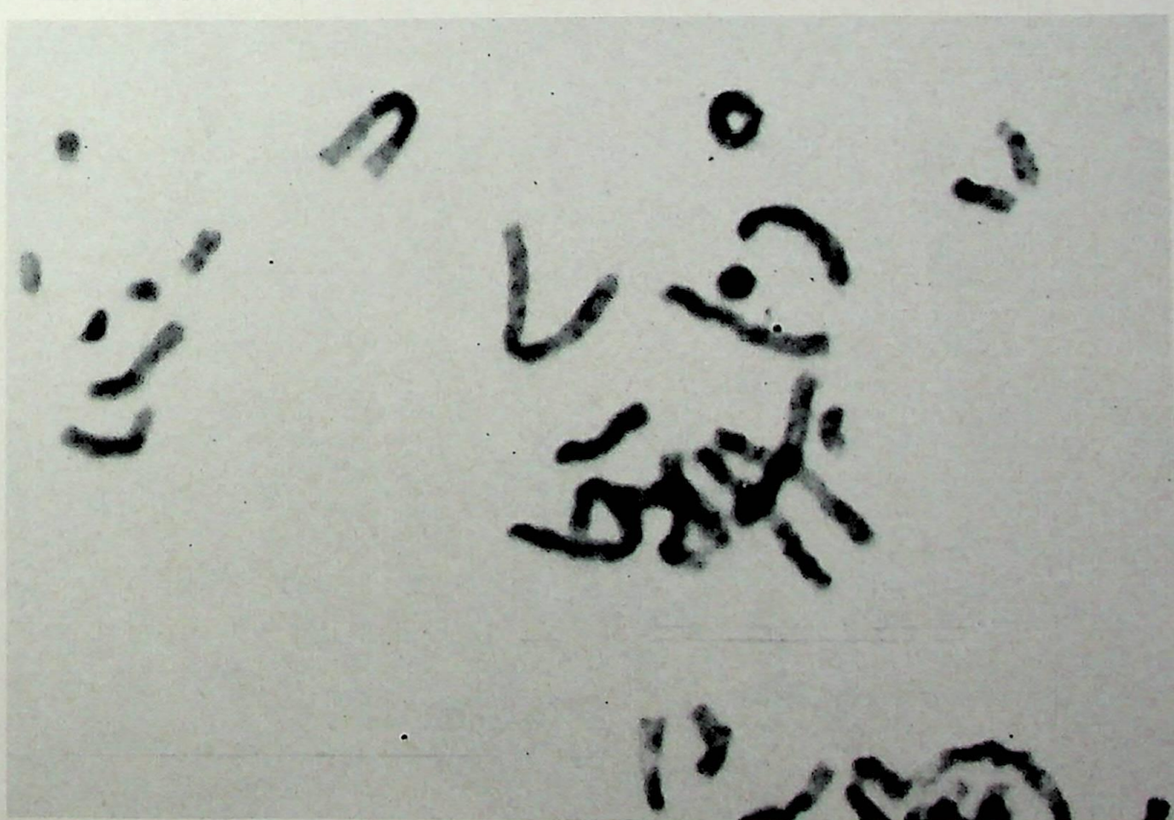
(a)



(b)

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

(c)



(d)

FIG. 9. Typical kinds of chromosome abnormalities observed when euploid human cells are irradiated *in vitro* with 230-kv x-rays. (a) Unirradiated cell. (b) Cell irradiated with 50 r. Breaks and deletions appear as indicated by arrows. (c) Chromosomes of cell irradiated with 75r, showing a translocation in the center, and a dicentric. (d) Portion of a mitotic figure of a cell after 150-r irradiation, demonstrating formation of ring chromosomes.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

one can safely conclude that the dose needed to introduce an average of one visible chromosome break per cell is no more than 40 to 60 r, a value which agrees within the limits of experimental uncertainty with that obtained as the mean lethal dose for colony formation of these cells.

The conclusion may be drawn that the genome of the euploid human cell is extraordinarily susceptible to chromosome damage by ionizing radiation, and that the average dose needed to introduce one chromosome break per cell is similar to the mean lethal dose to the cell's reproductive function, i.e., approximately 50 r. Since this value is far less than that required for the inactivation of various enzymatic activities in mammalian cells,[26] it appears to establish the fact that the cellular genome of man is far more sensitive to radiation damage which is, at least in part, presumably irreversible, than are other of its structures. While studies are currently in progress to determine whether cells *in vivo* will behave in the same way, experiments with radioprotective agents added to the medium suggest that only relatively small changes in the cellular D_0 value can be accomplished by changes in the cellular environment.

These considerations afford an understanding of the action of ionizing radiation on mammalian cells in terms of a primary effect—the initiation of a chromosome break, which requires on the average an exposure of a normal human cell to 50 r or less, and a secondary effect—the production of complex chromosomal aberrations through abnormal restitution of the fragments resulting from multiple breaks occurring within the same cell. The first of these involves a variety of alternative possibilities: The cell may continue reproduction with no recognizable cytogenetic change, though with perhaps a gene mutation at the site of the original chromosome break; it may suffer loss of all or part of a chromosome which could impair its ability to reproduce; or it may undergo formation of a chromosome bridge at anaphase which will certainly end its reproductive ability. When multiple chromosome fragmentations leading to complex translocations occur, the cell is much more likely to be incapacitated with respect to reproduction. Different cells should display 1-hit, multiple-hit, or intermediate types of survival curves, depending on the degree to which these different processes operate in causing destruction of the ability to multiply.

This analysis appears to offer an explanation for the difference in radiobiologic response exhibited by various mammalian cells. For example, normal diploid cells of different species, differing in their total mass of chromosome material, should be expected to exhibit differences in radiation sensitivity roughly though not exactly paralleling their chromosomal volumes, since the cell with greater chromosomal mass will offer greater opportunity for a given radiation dose to produce ionization directly or indirectly effective within this region. In accordance with this expectation, we have found that cells of the chick, whose DNA content is less than one-quarter that of human cells, possess a mean lethal dose for x-rays approximately 4 to 8 times greater than that of euploid cells of man. In Table II is presented a comparison of the D_0 values and DNA content of diploid cells from a variety of different species. Despite relatively large uncertainties in the individual determinations of the mean lethal dose and of the DNA content, and the large range of values encompassed, it is evident that these very different diploid cells exhibit a remarkably constant value for the product of these 2 variables. Other aspects of these relationships are considered elsewhere.

These considerations afford insight into the effects of various chromosome conditions like polyploidy on the radiosensitivity of mammalian cells from the same species. With increasing numbers of chromosome sets within a cell, its sensitivity to killing by a 1-hit process should decrease. Thus, a haploid cell will lose its ability for indefinite multiplication as a result of damage to any single gene essential to replication. Diploid cells which have a double gene set presumably will require loss or aberration of larger portions of a chromosome before the genetic imbalance necessary for inhibition of multiplication results. Since such chromosome losses readily occur as a result of 1-hit processes following exposure to ionizing radiations, such cells will often exhibit curves with a hit number close to unity, but may in some cases be multiple hit, if the gene distribution among the chromosomes is such that death rarely follows a single-hit process. The D_0 value of such curves will be influenced by the volume of each chromosome, the magnitude of the breaks which are produced, and the degree to which 1-hit lethal and 2-hit lethal processes occur when cells are exposed to various types of irradiation under various conditions. Cells with higher degrees of ploidy will be much less susceptible to reproductive killing by loss of part or all of any chromosome because of the smaller imbalance produced when larger multiples of the chromosome set are present. In such cells, the major lethal process will involve interaction between

TABLE II. Comparison of radiation and the DNA content for diploid cells of three different living forms. The data indicate that, despite individual variations in D_0 values and DNA contents of over a hundred-fold, the product of these two variables remains reasonably constant, varying only by a factor of about two.

Diploid cell type	D_0 value[a]	DNA content (in picograms per cell)	Product: $D_0 \times$DNA content (roentgens $\times$picograms)
Yeast	about 8000 r[b]	0.05[b]	400
Chick	about 300–400 r[c]	2.5[d]	900
Man	50–60 r[e]	8.3[d]	500

[a] The D_0 value is taken as the dose needed to reduce the reproducing fraction of the cell population to 37%, in the region where the survival curve is linearly exponential.
[b] See reference 27.
[c] See reference 25.
[d] See reference 28.
[e] See reference 32.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

two or more chromosomes to produce malformations like anaphase bridges. Hence, such cells will usually display a hit number in the neighborhood of 2. While the larger volume of the chromosome set in such cells might at first be considered to make them more vulnerable than diploid cells to the accumulation of hits that can lead to cell death, this factor may be more than counterbalanced by the relatively low probability with which two simultaneously broken chromosomes will restitute abnormally, so as to form a nondividing configuration (like a dicentric) as opposed to a union still permitting normal mitosis. Thus, the HeLa S3 cell, an aneuploid human malignant cell with 78 chromosomes, exhibits a survival curve which is approximately 2-hit, as opposed to the more nearly 1-hit curve of the euploid cell, and exhibits a D_0 value of 96 r instead of 50 r characteristic of the normal diploid cell. These parameters—the degree of cell ploidy; the distribution among the different chromosomes of regions whose loss may lead to cell death through imbalance; and the separate probability of formation of chromosomal aberrations which mechanically prevent unlimited division by 1-hit and 2-hit processes distributed among the various chromosomes—would appear to afford explanation of the difference in radiosensitivities of different cells. Quantitative evaluation of these parameters in selected normal and malignant cell systems is now under study.

These considerations also can offer explanation for our recent findings that the S3 cell which is enormously more sensitive than bacterial cells to killing by x-rays, is much more resistant to ultraviolet irradiation. Ultraviolet, while a highly effective lethal and mutagenic agent for bacteria, is known to be much less efficient than x-rays in producing chromosome breaks. While haploid bacterial cells can be killed effectively by point mutations and so are readily inactivated by ultraviolet light, the need for chromosome breakage to occur before the multiploid animal cell is inactivated would render it resistant to the action of ultraviolet irradiation.

Application of these aspects of the action of ionizing radiation at the cellular level, to analysis of the underlying mechanism involved in the pathologies arising from acute, whole body irradiation of animals leaves little doubt that many of the most important features of the mammalian radiation syndrome are due to the cumulative actions on the individual body cells of the same kind as those described in the preceding paragraphs:

(a) Knowledge of the D_0 value for individual cells for the first time provides clear explanation why the mean lethal dose for total body irradiation of mammals should lie in the range of 400 to 600 r. Exposure of the whole body to a dose of this magnitude would inactivate the reproductive mechanism of more than 99% of the cells with radiosensitivities like those of the cells studied *in vitro*. The body might recover from smaller doses which would still leave sufficient intact cells which could reproduce at a rate rapid enough to compensate for the losses. The mean lethal dose to the whole animal represents a point of loss of reproducing cells so extensive as to provide widespread damage to the integrities of physiological function which depends on cell reproduction and which constitutes a stress which cannot ordinarily be withstood.

(b) This formulation is in accord with the failure to find any specific biochemical lesion resulting from animal irradiation in the mean lethal range. Significant depression of individual enzyme functions have been found to require much higher doses than that needed to kill the whole animal. While no specific enzyme function could be implicated, many studies showed that DNA synthesis was greatly reduced as a result of irradiation in the dose range here considered.[29] The theory here considered would predict such behavior. DNA synthesis can proceed normally only when cell reproduction progresses. Destruction of this function through production of chromosome aberrations must eventually depress DNA synthesis, even though no enzyme system of the cells or body fluids has been significantly altered by the irradiation.

(c) Similarly, the failure to find any accumulation of toxic material in the body fluids after exposure to radiation in the mean lethal range is to be expected. While at higher doses significant amounts of toxic products might be produced as a result of chemical structural alterations resulting from ionizations produced in the tissue, doses of several hundred roentgens only will alter significantly structures with a size of the order of magnitude of the chromosomes, and will produce biological changes only if the uninterrupted integrity of such structures is vital to normal body operations as is indeed the case with the genetic apparatus of the individual cells.

(d) It follows that one would expect those tissues to be most radiosensitive which must maintain the highest rate of cell mitosis, since these will suffer the most immediate loss of their specific functions, on which the body depends. This correlation between radiosensitivity and mitotic rate has been recognized as one of the earliest generalizations to emerge from radiobiologic experience. However, while the rapidly dividing tissues are indeed most radiosensitive from the point of view of their immediate contribution to embarrassment of the whole organism through failure to maintain their specific functions, it is necessary to regard virtually all of the body cells as equally sensitive to chromosomal damage. Each normal cell nucleus has an equal probability of accumulating similar chromosomal injuries. Such aberrations will remain largely latent in each cell until it comes to reproduce. Hence, the cells of the more rapidly dividing tissues are simply the first to make evident the injuries which must be regarded as equally numerous in the more slowly multiplying

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

regions of the body. This fact is of vital importance in understanding genetic transmission of radiation damage, and the ability for tumors to develop many years after a radiation experience.[30]

(e) Similarly, our picture accounts in straightforward fashion for the characteristic lag period between radiation exposure and the development of symptoms, which has long been recognized as typifying mammalian injury by ionizing radiation. The cell which has suffered damage which is primarily only chromosomal, will continue to function effectively for a period, since its full complement of enzymes and other structures presumably remains largely unaltered. Not until the cell reaches the point where it should normally divide will it begin to exhibit grossly deviant behavior from the normal. Moreover, we have shown that such cells may even multiply for several generations, before they and their progeny cease reproduction.[19]

(f) The sequence of events which ultimately leads to death of any acutely irradiated animal may thus be explained as follows. Chromosomal aberrations introduced among all of the cells exposed will first reveal themselves in those cells of the most actively mitotic tissues. Division will be prevented in most of such cells irradiated with several hundred roentgens, with the result that the functions supplied by these tissues will fail and the body will be threatened. As other more-slowly dividing tissues reach the point where they too should produce cell proliferation, they will add the results of their failure to those which have already introduced distortion of normal body physiology. In contrast to the actively dividing structures like the bone marrow, and the epithelial linings of the gastrointestinal tract, tissues like nerve and muscle in which cell division rarely occurs, can continue to exhibit normal function even after exposure to many thousands of roentgens. It is of interest that cell reproductive death has also been identified as the major factor in the mammalian radiation syndrome by Quastler and his co-workers, using experimental approaches and techniques completely different from those employed here.[31]

In this connection, an alternative hypothesis has been proposed that the more-rapidly dividing tissues are more radiosensitive only because cells in mitosis are in a more sensitive condition and hence are damaged more readily by irradiation. This proposal does not fit the facts, since even in the most rapidly dividing tissues, no more than 3% of the cells are in mitosis at any one time. Hence, if this were the true explanation, it could account at most for a negligible reduction in cell viability of such tissues. However, in early embryonic development, where high mitotic frequencies often are achieved and where phased cell reproduction may occur to bring many cells into mitosis simultaneously, this factor might easily play an important role.

(g) These considerations also explain quite naturally how it is possible to save animals irradiated with doses in the mean lethal range, by injection of viable bone-marrow cells, though not by cellular fragments or extracts. Such cells recolonize the irradiated tissue which is being depleted through impaired cell-reproductive function. While other tissues have presumably suffered equal diminution in the percent of cells capable of reproduction, these require replenishment less rapidly and hence provide more time for the surviving cells to recolonize the tissue, provided the bone-marrow functions can be maintained. It may be expected that animals might be saved from lethal effects of even higher doses, by additional inoculation of viable cells from other tissues which, as the dose increases progressively, would become critical in their failure to maintain functional integrity of the body.

(h) Similarly, the interesting experiments first carried on by Patt and his co-workers,[32] in which animals were subjected to low temperatures immediately after irradiation, become explicable. Such animals failed to develop any of the symptoms of radiation injury until after their temperature was restored, after which the entire sequence of pathogenesis was initiated as though the irradiation had just occurred at the time their body temperatures had been raised back to normal. At the low temperature, mitosis is inhibited so that the body does not produce the conditions which can bring into expression the latent damage to the cellular genetic apparatus. On rewarming, normal cell reproduction is again initiated, so that each time a genetically damaged cell comes to mitosis its injury becomes functional and contributes to embarrassment of a normal function.

(i) One might expect, then, to find some rough correlation between the D_0 obtained for euploid cells of different animals as measured by the survival curves here described, and the mean lethal dose for the whole animal. While the latter figure must, of necessity, be influenced by many complex interactions of an unpredictable kind, it is of distinct interest to find that such a correlation is at least suggested by the small amount of currently available data, as shown in Table III.

Studies of the action of radioprotective agents on x-ray survival curves of S3 cells are completely consistent with the interpretations here developed. The compound, 2-mercaptoethyl guanidine, which has been demonstrated to raise the MLD for mice from 900 to a maximum of 1450 r,[35,36] also exercises significant radio-

TABLE III. Data demonstrating that possibly some parallelism may exist between the mean lethal dose for whole warm-blooded organisms (LD_{50}) and the D_0 values of their single, euploid cells as determined *in vitro*.

	X-ray LD_{50} for whole animal	X-ray D_0 of euploid cell *in vitro*
Man	400– 500 r	50– 60 r
Chinese hamster	825–1190 r[a]	160 r
Fowl	1000 r[b]	300–400 r

[a] See reference 33.
[b] See reference 34.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

protection on S3 HeLa cells plated *in vitro*. The shape of the survival curve remains approximately 2-hit as it was in the absence of the compound, but D_0, the mean lethal dose for cell reproduction, is raised from 96 r to a maximum of 160 r, a value which agrees far better than perhaps could be expected with the degree of protection achieved in the whole animal. All of the data on the kinetics of the radioprotective action of this compound on single plated cells are consistent with the interpretation that the presence of this material lowers the effective dose of x-rays which reaches the sites within the cell whose inactivation results in loss of the ability to reproduce indefinitely.

(j) Finally, these data support the cellular-genetic interpretation of the great effectiveness of relatively low doses of radiation in suppressing antibody formation in the mammalian body. The dose range needed to inhibit antibody production significantly lies in the region of 50 to 150 r,[37] a value which, through its close correspondence with the mean lethal dose for euploid cell reproduction, suggests this action to reflect the effect of radiation in preventing multiplication of antibody-producing cells. By contrast, specific macromolecular biosynthetic mechanisms appear essentially undamaged even after irradiation of mammalian cells with thousands of roentgens.[15] Thus, the great radiosensitivity of antibody production suggests that the mechanism of new antibody production involves the need for cell multiplication, rather than simply antibody synthesis by pre-existing cells.

The fact that some mammalian cells exhibit survival curves for the reproductive function which are multiple-hit in character must not be taken as an indication that a threshold exists for the induction of mutations by high-energy radiations. On the contrary, all of the evidence here presented leads to the conclusion that, while a cell displaying a 2-hit survival curve does display a threshold for killing by high-energy radiation, the production of single chromosome breaks, and the attendant alteration of genetic material at the site of such a break, goes on even when doses insufficient to kill are applied. From the point of view of the entire organism and its progeny, it may be far better for any cell to be permanently inactivated by multiple chromosomal aberrations than to suffer only a single chromosomal change, which then may be handed on to the offspring, usually as a recessive genetic defect.

The data here presented make possible an estimation as to whether background radiation may contribute to the processes of aging in man. Since 50 r is the average dose needed to produce 1 visible chromosome break per cell by the techniques here employed, one may well expect that damage on a smaller scale which cannot be seen by ordinary microscopy results from even smaller doses. Thus, one can calculate that, with a background exposure of 0.1 r per year, in 70 years the accumulated dose is sufficient to have produced gross chromosome damage in one-seventh of all of the body cells, and probably more subtle effects in a much larger number. This would appear not to be a negligible process, and thus makes it probable that the gradual attrition of body functions which constitute aging, may be, to a very significant degree, the reflection of accumulation of cellular genetic injuries in cells and their descendants originating from background irradiation. By extension of the methodologies here discussed, it is proposed to attempt analysis of physiologic and genetic differences in the behavior of cells taken from aging animals, and to examine as well the effects of the molecular constituents of body fluids from such subjects on physiologic and genetic behavior of standard cell strains *in vitro*.

Discussion of the use of quantitative methods in measuring mammalian cell growth and genetics would be incomplete without at least mention of the outstanding development by Dulbecco and his associates[38] in achieving a plaque technique for precise enumeration of single particles of mammalian viruses exactly as has been current in bacteriophage studies. Thus, it becomes possible also to quantitate virus and cell interaction in animal systems, so that one may expect equally profound insights to arise from such studies as has been the case in the interaction of bacteriophages with their bacterial hosts.

The thesis which has been the purpose of this discussion is that newer developments promise to provide well-defined systems for quantitative exploration of physical and physicochemical mechanisms involved in growth, genetics, and differentiation in mammalian-cell systems. There is every indication that in the space of a few years it will be possible to come to grips with specific molecular aspects of these basic mechanisms in as intimate a fashion as is now current in the biophysical approach to microbial systems.

BIBLIOGRAPHY

[1] T. T. Puck, P. I. Marcus, and S. J. Cieciura, J. Exptl. Med. **103**, 273 (1956).

[2] T. T. Puck and H. W. Fisher, J. Exptl. Med. **104**, 427 (1956).

[3] T. T. Puck, J. Cellular Comp. Physiol. (to be published).

[4] H. Eagle, J. Biol. Chem. **214**, 839 (1955).

[5] V. J. Evans, J. C. Bryant, W. T. McQuilkin, M. C. Fioramonti, K. K. Sanford, B. B. Westfall, and W. R. Earle, Cancer Research **16**, 87 (1956).

[6] J. F. Morgan, H. J. Morton, and R. C. Parker, Proc. Soc. Exptl. Biol. Med. **73**, 1 (1950).

[7] H. W. Fisher, T. T. Puck, and G. Sato, Proc. Natl. Acad. Sci. U. S. **44**, 4 (1958).

[8] I. Lieberman and P. Ove, J. Biol. Chem. **233**, 637 (1958).

[9] F. C. Parker, L. N. Castor, and E. A. McCulloch, "Cellular biology, nucleic acids and viruses," special publication of New York Acad. Sci. **5**, 305 (1957).

[10] A. E. Moore, C. M. Southam, and S. Sternberg, Science **124**, 127 (1956).

[11] A. A. Axelrad and E. A. McCulloch, Stain Technol. **33**, 67 (1958).

[12] K. H. Rothfels and L. Siminovitch, Stain Technol. **33**, 73 (1958).

[13] J. H. Tjio and T. T. Puck, J. Exptl. Med. **108**, 259 (1958).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

[14] T. T. Puck, S. J. Cieciura, and A. Robinson, J. Exptl. Med. **108**, 945 (1958).
[15] J. H. Tjio and T. T. Puck, Proc. Natl. Acad. Sci. U. S. (to be published).
[16] J. H. Tjio and A. Levan, Hereditas **42**, 1 (1956).
[17] G. O. Gey, W. D. Coffman, and M. T. Kubicek, Cancer Research **12**, 264 (1952).
[18] J. H. Muller in *Radiation Biology*, A. Hollaender, editor (McGraw-Hill Book Company, Inc., New York, 1954), p. 351.
[19] T. T. Puck and P. I. Marcus, J. Exptl. Med. **103**, 653 (1956).
[20] C. M. Pomerat, Ann. New York Acad. Sci. **71**, 1143 (1956).
[21] T. H. Ingalls, H. R. Morrison, and L. I. Robbins, New Engl. J. Med. **2581**, 252 (1958).
[22] T. T. Puck, D. Morkovin, P. I. Marcus, and S. J. Cieciura, J. Exptl. Med. **106**, 485 (1957).
[23] E. K. Y. Chu and N. M. Giles, J. Natl. Cancer Inst. **20**, 383 (1958).
[24] M. A. Bender, Science **126**, 974 (1957).
[25] T. T. Puck, Proc. Natl. Acad. Sci. U. S. **44**, 772 (1958).
[26] Z. M. Bacq and P. Alexander, *Fundamentals of Radiobiology* (Butterworths Scientific Publications, London, 1955), pp. 228–262.
[27] C. A. Tobias, Natl. Acad. Sci.-Natl. Research Council, Publ. No. **18**, 46 (1956).
[28] I. Leslie in *The Nucleic Acids*, E. Chargaff and J. N. Davidson, editors (Academic Press, Inc., New York, 1955), Vol. II, p. 1.
[29] Z. M. Bacq and P. Alexander, see reference 27, pp. 257–258.
[30] C. L. Simpson and L. H. Hempelmann, Cancer **10**, 42 (1957).
[31] H. Quastler, Radiation Research **4**, 303 (1956).
[32] H. M. Patt and M. N. Swift, Am. J. Physiol. **155**, 388 (1948).
[33] G. Yerganian, Federation Proc. **14**, 1371 (1955).
[34] Z. M. Bacq and P. Alexander, see reference 27, p. 220.
[35] D. G. Doherty (personal communication).
[36] R. Shapira, D. G. Doherty, and W. T. Burnet, Radiation Research **7**, 22 (1957).
[37] H. T. Kohn, J. Immunol. **66**, 525 (1951).
[38] R. Dulbecco and M. Vogt, J. Exptl. Med. **99**, 167 (1954).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

48
Interactions between Cells

PAUL WEISS
The Rockefeller Institute, New York 21, New York

INTERACTIONS among cells are the means by which the cell community of the organism establishes and maintains its organizational harmony. They are so numerous and varied that it would take more space than is allotted here just to draw up a reasonably comprehensive list. Yet, knowledge about them is still very scanty. Fascinating progress has been made in the study of some of the biochemical and biophysical components of the living system. But knowledge seems to have grown the faster, the smaller the sample of the living system that was taken under investigation. The major advances were made down at the molecular level. The task of dealing with the larger cellular level, therefore, involves a re-entry into areas of uncertainty, and the course from the introductory article on cell dynamics (Weiss, p. 11) to the present article marks a full circle from relative ignorance through knowledge back to relative ignorance. The study of cells in interaction deflates complacent notions that all the major facets of cell life are truly understood even in principle.

"Interactions between cells" covers practically everything that is going on in organisms, and obviously this account must confine itself to a few crucial examples. There are essentially two ways for cells to interact: either a cell elaborates a diffusible substance which affects another cell at some distance, or a cell transmits an effect to another cell by direct contact. Since mediation by diffusible agents, as in the hormone system, is more widely studied and better understood than are the contact interactions, the emphasis here is on the latter. A more detailed account can be found in the author's recent review article on "Cell Contact."[1]

First, a few words on what is meant by "contact." It has been very easy to define contact between cells in terms of observations with the ordinary light microscope. If microscopically two cell borders came so close as to merge into a single line, there was contact, whereas a microscopic space in between signified lack of contact. Conversely, any interruption of the microscopic outline between two cells was interpreted as "protoplasmic continuity." But evidently, such questions of "contiguity" *vs* "continuity" between cells are merely questions of the resolving power of the instruments at hand. It is not surprising, therefore, that the introduction of electron microscopy has brought a marked change of outlook.

In the first place, electron microscopy has revealed a higher degree of organization at cell surfaces than previously assumed. As an example, one can cite the larval amphibian skin referred to in the introductory article (Weiss, p. 11). The underside of the epidermal cells, which rest on the basement membrane, is dotted at regular intervals with submicroscopic bodies, about 1200 A high, consisting each of two electron-dense round plates connected by a lighter neck.[2] The outer one of the plates forms part of the cell surface. This surface itself is separated from the underlying basement membrane by a gasket-like granulated film of a few hundred Ångströms thickness. The cell surface thus is actually a mosaic of patches of supramolecular order and different chemical and physical properties, which have been partly identified. This was done by applying various enzymes to fragments of live skin before they were fixed in osmium tetroxide and prepared for electron microscopy. In pancreatic lipase, the dark plates lose their osmiophilia, indicating that normally they have a high lipid content. In distilled water, the neck of these bodies swells and breaks, leaving unconnected double plates. Thus, the neck region is hydrophilic. Salivary amylase dissolves the film through which the epidermal cell is attached to the basement lamella, suggesting abundance of carbohydrate. It also erodes the parts of the cell surface between the dense bodies, but leaves the latter intact and protruding, thus identifying them as solid bodies.

Such observations demonstrate that it would be quite mistaken to consider a cell as being equal and uniform over its whole surface. Characteristically, this surface pattern is confined sharply to the portion of the cell that is in immediate apposition to the basement membrane, thus indicating a direct contact interaction between cell and substratum. This supposition is confirmed by experiments in which the contact between the two structures was first broken and then restored, as follows. When part of the skin is injured, the epidermal cells near the wound roll off. The dense bodies, which seem to serve as suckers to attach the cells to their substratum, become detached and are resorbed within a few days. They are re-formed during wound healing as the epidermal cells migrate over the defect, but are precisely confined to that fraction of cell surface now in fresh contact with the substratum. In other words, the cell has a mechanism for producing this specialized apparatus which responds to the interaction between cell surface and the underlying carbohydrate-rich film. As a result, an "induction" on the submicroscopic level occurs. The cell is induced along a geometrically and physically defined surface to display a specific fraction of its synthetic repertory. Other instances of contact induction are given later in this paper.

When two cells are in contact, the problem becomes more complicated. At the contact points between two

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

epidermal cells, single dark plates are present, the well-known nodes of Bizzozero. Electronmicroscopically, they show as a single dark plate in each of the contacting cells with a lighter space in between.[1] Odland[3] has recently scanned this region densitometrically in superior electron micrographs and has found an additional dark line in the center of the light interspace. This would indicate that the molecular array separating the two cell surfaces is not random, but has some degree of spatial order. It may be that the cell surfaces are linked by macromolecules normal to the surfaces of some 100 A in length, which is the order of width of the light space. Gaps of similar dimensions have been found between practically all cells that make intimate connections with each other, including synapses between nerve cells. It would seem best, therefore, to define two cells as being in contact, operationally speaking, when they are separated by a space whose contents are not subject to random perturbation. Depending on the dimensions, several factors might operate to limit perturbation of the interspace. For example, macromolecular compounds moving out of the cells may form bridges between organized parts and establish special submicroscopic attachments or interaction points between the adjoining cell surfaces. Evidence from tissue culture, obtained over a decade ago,[4] indicates that certain cells, and perhaps all cells, when left to their own devices, surround themselves with characteristic colloidal exudates, which must be taken into account when cell-to-cell relations are analyzed. Such surface coronas, extending for unknown distances into the cellular environment, might be major factors in immune reactions, cellular aggregations into colonies, phagocytosis of specific substances, selective associations among nerve fibers, and the like. It is equally possible that cell surfaces react with each other specifically without such mediators. A few examples are cited in the following.

In experiments on the development of nerve fibers, it has been found that fibers of the same type have a selective affinity to one another—motor fibers applying themselves to motor fibers, and sensory fibers to sensory fibers, and within each class, each subgroup to its corresponding type.[5] Only by virtue of such specific grouping is it possible for a surgeon, for instance, to expose the spinal cord and sever discriminately a bundle of pain fibers or of fibers for deep perception. Such fibers would not run in common bundles unless they had grown out that way, and since they do not develop all at the same time, the older fibers must have served as guides to latecomers of the same character.

The same principle applies, even more subtly, to the relations among the cells within the central nervous system on which the coordination of neural functions depends. Our current concern with the nervous system as a system centers mostly around such matters as geometric properties, electric parameters and time constants—length of interconnections within the network, number and distribution of branches, temporal characteristics of the individual units, synaptic resistances, etc. We generally ignore the fact that there is great biochemical diversity within the system far beyond the gross distinctions of cholinergic and adrenergic portions, which makes it work as something more than a monotonic network of conducting fibers. Developing nerve cells have highly selective discriminatory affinities by which they "recognize" each other and their surroundings. This introduction of the anthropomorphic term of "recognition" is not anything one need apologize for so long as physicists speak of electrons "seeing" each other. Recognition of one cell by another may be based on conforming charge distributions, on conforming molecular groupings, or on as yet wholly unsuspected mechanisms. The problem of explaining such recognition is of course encountered in many other biological phenomena, e.g., in enzyme-substrate relations and in antibody-antigen reactions. With a few notable exceptions,[6] affinity reactions among somatic cells, however, have received little attention and even less methodical study, even though there is ample evidence for the widespread occurrence of such selective behavior.

To cite a specific example, cells can "locate" their proper destinations in the body even if they are deprived of their customary routes for getting there. This has been shown[7] by letting embryonic cells of known destination be distributed at random through the blood stream of a host embryo. In order to be able to identify the injected cells, one chooses cells which carry a marker. We chose the precursor cells of pigment cells which, when introduced into nonpigmented breeds of chicks, would reveal their origin by synthesizing black melanin. It was found that after an injection of such cells into the blood stream, about 8% of the host chicks later carried scattered patches of black pigment cells in their feathers and skin. These pigment cells were always situated in the precise positions where such cells would have resided in the donor animals, and in no case was a pigment cell ever observed to have become lodged where pigment cells do not normally belong. One must conclude, therefore, that donor propigment cells had colonized only specified areas in the host embryos where such cells normally end up, which, in view of the random dissemination of the dissociated cells in the host body, implies the existence of a mechanism for highly selective localization. As few as one or two propigment cells, when arriving by chance at an appropriate site, "recognize" it and settle down to proliferate and differentiate into pigment cells. Cells that miss encountering within the embryo an environment favorable to their type fail to become lodged and to differentiate, and presumably are resorbed.

On the other hand, outside the embryo, on the yolk sac, where specific sites for embryonic cells are lacking, clumps of injected cells get stuck, and their fate led to a further significant extension of these observations. It was discovered[8] that, where random mixtures of em-

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

bryonic cells from a variety of sources (nerve, muscle, cartilage, glands, etc.) had become lodged on the yolk sac, they gave rise not to indiscriminately mixed structures, as might have been expected, but to quite harmoniously organized organ complexes. Some degree of self-sorting according to types must have taken place among these scrambled cells after they had become trapped.

Further conclusive evidence of "self-sorting" was found in tissue cultures in which random mixtures of suspensions of cells of different embryonic origins —for instance, cartilage and kidney cells—had been combined.[9] These cultures developed compound structures in which islands of pure cartilage were sharply demarcated from blocks of pure kidney tissue, indicating that cells had become re-associated like-to-like. Thus, even *in vitro* cells, which have acquired in their prior embryonic history some biochemical differential related to their subsequent development as either kidney cells or cartilage cells, can assort themselves according to type. Since the cells make contact only with their surfaces, one must assume that some property associated with differentiation into kidney cells or into cartilage cells has imparted type-specific markers to their surfaces for mutual recognition. But how such "recognition" leads to active sorting remains obscure.

How do mixed cell populations achieve this orderly reassortment? Do like cells "attract" each other? Or do they "recognize" each other only after chance encounters? A clue came first from some observations on selective wound healing after the grafting of various kinds of epithelia into skin gaps in the flanks of amphibian embryos.[10] The first coverage of a wound is effected by the migration of epithelial cells of the wound edge over the raw surface. When these advancing cells reach a graft, fusion of the two fronts occurs only if the graft contains an epithelium of a type which normally borders on skin epidermis. In that event, the advancing cells stop in their tracks, so to speak, merge into a uniform sheet with the graft, and no further proliferation occurs. Implants of skin, or of cornea or oral lining—which are normally continuous with skin—are accepted in this manner. But, when a fragment of lung or gall bladder or esophagus is implanted in a skin wound, the migrating skin epidermis is not halted on contact with those foreign-type cells. The epidermal edges glide over or under the graft and continue until they meet with the skin advancing from the opposite side.

Results of this kind have long been known in surgery and lead to the conclusion that, whether we have a plausible explanation for the phenomenon or not, cells of identical or conforming constitution (as epidermis, cornea, and oral lining) will recognize each other as akin and stay put; whereas cells of unlike character develop a reaction which will cause them to separate.

With these facts as background, we proceeded to study the contact reactions between different kinds of cells directly *in vitro*,[1] by taking phase-contrast time-lapse motion pictures of encounters among cultured cells.* If the cells that make contact are all of the same kind—all liver cells, for example—they aggregate and stay together. The free borders of roaming cells are in constant motion; but wherever two like cells touch each other, that portion of their surfaces becomes quiescent. The bond between them, however, is not static; there is no cement which sticks them together as in an immunological precipitin reaction. On the contrary, the cells continually glide over one another, constantly changing their positions relative to each other and, by the time others have joined them, their positions in the group.

On the other hand, if cells of two different types are cultured together (e.g., lung and liver or lung and kidney), the result is quite different. As the loose cells stray about, they collide at random. There is not the slightest indication that they would react to each other's presence unless or until their free borders touch. Neither do like cells "attract" each other, nor do unlike cells avoid each other. Mutual recognition and consequent discriminatory behavior are decidedly contact responses. All cells, whether alike or not, make primary contact indiscriminately. But, whereas like pairs thereafter draw closer together and remain joined, unlike pairs separate secondarily by reciprocal withdrawal of those parts of their borders that had been in fleeting touch. While the mechanism of identification undoubtedly resides in the surface, the interior of the cell seems to be engaged in the withdrawal reaction, giving the phenomenon the general appearance of a "reflex" response with a "sensory" and a "motor" component. The "reflex time" is of the order of 10^3 sec, varying both with the strangeness of the confronted cells and with their respective rates of motility. In conclusion, self-sorting of scrambled cell types results from chance collisions with matching combinations holding on to each other, whereas nonmatching ones do not last.

"Matching" combinations involve either cells of the same type or of complementary types, the latter being types that normally cooperate in the building and functional operation of complex organs. For instance, an active mutual adhesion between nerve processes and enveloping sheath cells has been demonstrated,[11] and our motion pictures reveal a similar marked tendency of macrophages to confine their excursions to within the borders of the large flattened lung cells with which they have been jointly explanted.

The discriminatory response is strictly cell-type specific, but it is not species specific. That is to say, cartilage cells of the chick will reject association with liver or kidney cells of the same chick, but they will combine readily with cartilage cells of a mouse. This fact, derived earlier from reaggregation experiments,[12] has now

* These moving pictures, made in collaboration with A. C. Taylor and A. Bock, were shown at the Study Program in Biophysical Science in conjunction with the lecture on which this article is based.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

been proven visually by the motion pictures. On the other hand, the same cell type may undergo progressive alterations with age, so as to make more mature and less mature cells of the same strain less acceptable to each other than cells of the same age would be. Furthermore, there are signs that the differentials among cell types, which underly their discriminatory responses to each other, may be subject to gradations in accordance with their ontogenetic relationships.

The details of this principle still remain to be worked out. But what has emerged thus far is enough to force a substantial revision of former excessively static pictures of the organism as a fixed framework of parts shifted into predestined places by a rigidly prescribed system of tracks and schedules and then immured by fibrous cements. Although, grossly, this picture still holds, it gains much greater flexibility in detail from the realization that the eventual pattern of combination, association, and segregation among cell types in typically ordered arrays is not just passively arrived at, but is actively insured and guarded—and restored after disturbance—by a system of subtle mutual conformances and nonconformances with which the various cell types have been endowed as a corollary of their ontogenetic differentiation.

The principle of the self-linking of cells into matched groups by contact affinities and disaffinities has an important bearing on the explanation of the establishment and maintenance of order in the networks of the nervous system. The building up of central and peripheral nerve cables by the selective attachment of nerve fibers of a given type to others of like specificity has already been referred to in the foregoing. The same principle holds for interneuronal relations, as well as for the relations between neurons and their non-neural receptor and effector organs, on an even subtler scale. Since this is one of the most compelling, yet at the same time least recognized, instances of specificity and selectivity in cell-to-cell interactions, it is well to recapitulate briefly the relevant experiments, which I started almost forty years ago (see a recent review[13]) and which have more recently been amplified by my student, Sperry.[14]

During development, countless connections are established among individual neurons and between neurons and peripheral sense organs and muscles. The latitude inherent in the primary developmental mechanisms of neurogenesis[5] is so great that it would rule out any such microprecision in the details of the neuronal circuitry as is usually postulated as basis for coordinated functions. On the other hand, there is equally conclusive evidence to show that the nervous system does emerge from its embryonic phase with a large list of ready-made coordinated performances, which, since learning by trial and error could have played no part in their formation, must be explicable in terms of developmental interactions. The dilemma was solved by the demonstration of secondary processes which adjust the details within the gross primary system of connections to the unique arrangements of each individual specimen. Although nerves from the same central sources may in different individuals end up in different muscles, the muscles themselves then send a specifying influence back into the centers with the information on just what muscle of what name lies at the end of what line.

The test experiment consisted of transplanting a supernumerary muscle (or a whole set of limb muscles) near a normal limb and hitching it to a random branch of the local nerve supply. This amounted to "tapping" the communication network by inserting an extra receiver. What was observed then was that the transplanted muscle responded in the course of normal activities of the animal always at the precise time and with the same strength as the muscle bearing the same name in the normal limb. Thus, each muscle is called up, as it were, by the central system by its proper name, and if there are several muscles of the same name present in a common innervation district, they all respond simultaneously when the particular code is called. Since the test muscle was inserted arbitrarily, it is clear that it must have established its name-specific correspondences in the nerve centers secondarily. The same correspondence between periphery and centers was also confirmed for the sensory part of the system, first for proprioceptive[15] and tactile[16] sensations and later for the visual field.[14]

It all boils down to the following. Even though it is impossible to distinguish different skeletal muscles by present biochemical methods, the nervous system can tell them apart very well. Whatever the telling difference between muscles, it must be something that can make itself known in a retrograde manner over the motor nerve fiber in such a way that the central apparatus thereby gains exact information about its peripheral terminations. Motor fibers are initially blank and unspecified. They acquire specific identity only after they penetrate identified muscles at the periphery, and the same holds for sensory neurons in regard to their endings. Somehow, the acquired detailed specificity is then transmitted still further back from the primary to secondary neurons, and so on up the lines.

By this device, the neuronal population becomes progressively specialized into subunits whose member cells can henceforth regulate their own mutual relations by virtue of their acquired distinctive properties. Neurons of corresponding or complementary specificities thus become joined into cooperative systems. The proposition that this implies only linkage of a static morphological order all the way through the nerve centers[14] does not seem to go far enough if one is to explain either the aforementioned experimental results or the general problems of neural coordination.[13] But, if one adds the assumptions: that synaptic contact between neurons is merely an enabling, but not a decisive, condition for impulse transmission; further, that actual transmission requires specifically matched states of two apposed surfaces; and, finally, that the specificities

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

underlying this conformance are subject to modification by both peripheral and central influences; one comes closer to a satisfactory concept. The nervous system would then emerge as a vast system of resonance circuits, the elements of which would be linked by conforming molecular surface configurations, partly permanently, partly variably in response to changing central and peripheral influences.

From these remarks, it can readily be seen that compatibilities and incompatibilities along the boundaries between neurons are as significant as discriminatory devices as are the signs of surface "recognition" described earlier in this article for other cell types. Indeed, the direct demonstration of cell discrimination in the motion pictures should help to make the corresponding property of neurons acceptable, if not more palatable, to those who had hoped to be able to maintain their oversimplified faith in the essential identity of all neurons so long as they were faced only with the more indirect earlier evidence. In reopening the whole question of specificity in the nervous system, which of late has been lying dormant, our observations on cell encounters in tissue culture thus assume added significance, as they also open the way to a more direct practical study of the nature of the specificities concerned.

The reference to the progressive diversification within nerve-cell populations points logically to the more elementary question of how any primitive cell types, including neurons, acquire their distinguishing characteristics in the first place. To discuss the problem of divergent differentiation would go far beyond the scope of this article. There is one aspect, however, that fits into the context. Differentiation occurs essentially by dichotomies in the courses of transformation of cells with basically identical physical and chemical endowments.[17] Any cell strain is, by virtue of such endowments, initially capable of effecting a limited variety of qualitatively different reactions. Every step of differentiation implies that from that multiple repertory one definite course has become activated in some members of the group, and an alternative course in the other members, the two courses being mutually exclusive. Sooner or later, the different reactions lead to manifest diversity. Development is composed of long series of such steps. Present evidence is that each step has its own mode of triggering mechanism, and that, contrary to earlier illusions, there is no single master agent that could be held generally accountable for the various and successive dichotomous changes. Formally, however, one can distinguish two classes—one in which the differentiating activity plays entirely within the bounds of a group of similar cells, referable perhaps to differential interactions among the members of the group depending on their relative positions; and another in which the reaction of a given cell group is decisively influenced by an extraneous cell group.

This second class is usually termed "induction"; again, different steps of "inductive" influences need have little in common but the name. Some of them operate over distances, others only between adjacent tissues. Evidently, these latter raise the issue of whether true contact interactions are involved. The answer seems to be that in some instances and for certain steps of differentiation, intimate contact between the interacting cells is essential, whereas, in other cases, such direct contact is dispensable, the interaction being mediated by diffusible substances.

Some cases in which direct contact is necessary may be cited as examples. When devitalized (frozen-dried and rehydrated) cartilage is transplanted into the vicinity of a limb bone in amphibian larvae, the implant can induce the formation of new cartilage in contact with its surface.[1] Thus, competent living host cells build on cartilage to the dead cartilage model as bees would build on new honeycomb to an artificial wax honeycomb presented to them. This is a particularly interesting case, because the bodies of the cells in the implant cannot have been the inducing agents directly, as they are enclosed and insulated in the cartilaginous matrix; therefore, the inductive stimulus to surrounding host cells must have originated in the exposed surface of that matrix. Another set of experiments showed the same thing for bone.[1] A frozen-dried and rehydrated metatarsal bone of a rat, inserted into the leg of a host rat, induced there new bone, including new bone marrow, in contact with the old bone. That contact is essential has been confirmed from other sources.[18] Such observations support the conclusion that extracellular materials may play important roles in certain inductive interactions.[19]

The next case is of special interest to students of collagen. The stroma of the cornea consists of layers of collagen in a characteristic regular arrangement. When frozen-dried cornea was transplanted into a corneal defect in rabbits, the grafted stroma as such persisted for many months. In addition, however, it induced along its inner side the formation of several new layers of typical stroma.[20] Thus, the architecture of the dead stroma matrix has in some fashion been transmitted to the induced layers, so that they assumed the characteristic pattern of collagenous corneal lamellae. This result again proves that not only cells but cellular products likewise can influence other cells inductively.

The intimacy of submicroscopic cell contact, in the sense of the introductory remarks, which may be required for "contact inductions" has not yet been clarified. One of the few clues is the fact that, during the period of inductive interaction between eye cup and prospective lens epithelium, the attachment between the interacting cell layers is so firm as to resist mechanical separation.[21] But, it is still undecided whether the specific transmission across the cell borders is simply an orienting or sorting influence on pre-existing molecular populations[22] or whether it involves the actual passage of substance,[23] perhaps prepared by molecular reorienta-

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

tion.[1] There is a vast field here for exact studies on the relations between physical configuration and chemical activities. These relations are bound to remain in obscurity so long as one focuses attention on those cell-to-cell influences that are mediated not by contact, but by remote control through diffusible agents. The latter, although equally real, can tell only part of the story of cellular interaction.

In summing up, to state that all of the specific interactions among cells exemplified in this article are based on specific properties of the interacting cells is a truism. But, it explains why so little is known about the molecular phenomena governing cell-to-cell relations. There just has not been enough preoccupation with the molecular basis of cellular specificity. Unquestionably, cell "recognition" and selective response must eventually be reduced to terms of properties of molecules and molecular populations. But, so far, there has been very little tangible progress in that direction. There have been a few hypothetical suggestions to explain cell-to-cell conformances, none of them quite tenable in the light of recent observations. I had proposed[24,25] that some sort of steric fitting (e.g., by corresponding charge distributions) between complementary molecules might produce bonds or linkages between like cells, whereas unlike cells without conforming molecules in their surfaces to interlock could establish no connections. Yet, this supposition is now clearly contradicted by the fact that like cells, while remaining together, are not fastened but move continuously around each other. This leaves us for the time being without substantial explanation of the phenomena of selective aggregation as shown in the films.

The only thing that seems clear is that these phenomena point to the same general area of specificity that covers immune reactions,[26] enzyme-substrate interactions, pairing of chromosomes, fertilization, parasitic infection, and phagocytosis. The latter clearly involves selective uptake of particles of suitable molecular surface organization. The problem is no different from that met in cell-to-cell encounters, except for the great disproportion of size between the members of the pair in phagocytosis. Presumably, the selective ingestion of macromolecules belongs in here, too, with the discrepancy of size even more pronounced. The opposite extreme is found in a cell which spreads out selectively over a large body of proper substratum, where the cell is very small in comparison to the partner which it tries to engulf. So, from phagocytosis and the uptake of macromolecules, through cell-to-cell contacts with specific "recognitions," up to the problem of selective adhesion of a cell to its substratum, one deals with a continuous spectrum of problems of the same nature.

This very fact holds promise of progress, as advances in any one sector of that broad area will shed light on other sectors. On the other hand, cleverly getting around the acknowledgment of specificity in any one sector, whether by theoretical constructs or by unrealistic models, will not strip the other sectors of their aspect of specificity, and, hence, will not relieve the intellectual discomfort engendered by our inability to squeeze the broad subject of specificity into the limited conceptual framework which we have erected from the study of fragmentary vital phenomena lacking that aspect. Specificity, as a real and basic control mechanism of cell behavior, can no longer be relegated to a corner, but must be placed in a central position in cellular and molecular biology. The less that is known about it, the greater is the challenge. If there has been some dodging in the past, it was because of some concealed hope that the whole thing might yet in the end turn out to have been an illusion born of inadequate penetration. Exactly the opposite has happened. The more penetrating the analysis, the more cogent has the evidence of specificity become; witness the motion pictures of cell discrimination. There seems little doubt that once the reality and universality of the problem are generally acknowledged, and the promising techniques at our disposal for its disciplined study are recognized, progress can be rapid. But there is even less doubt that, if the existence of the problem continues to be widely ignored, or even denied, some of the major clues to the understanding of living systems will remain missing.

BIBLIOGRAPHY

[1] P. Weiss, Intern. Rev. Cytol. **7**, 391 (1958).
[2] P. Weiss and W. Ferris, Exptl. Cell Research **6**, 546 (1954).
[3] G. F. Odland, J. Biophys. Biochem. Cytol. **4**, 529 (1958).
[4] P. Weiss, J. Exptl. Zool. **100**, 353 (1945).
[5] P. Weiss in *Analysis of Development*, B. H. Willier, P. Weiss, and V. Hamburger, editors (Saunders Company, Philadelphia, 1955), p. 346.
[6] J. Holtfreter, Arch. exptl. Zellforsch. Gewebezücht. **23**, 169 (1939).
[7] P. Weiss and G. Andres, J. Exptl. Zool. **121**, 449 (1952).
[8] G. Andres, J. Exptl. Zool. **122**, 507 (1953).
[9] A. Moscona, Proc. Soc. Exptl. Biol. Med. **92**, 410 (1956).
[10] J. J. Chiakulas, J. Exptl. Zool. **121**, 383 (1952).
[11] M. Abercrombie, M. L. Johnson, and G. A. Thomas, Proc. Roy. Soc. (London) **B136**, 448 (1949).
[12] A. Moscona, Proc. Natl. Acad. Sci. U. S. **43**, 184 (1957).
[13] P. Weiss, Symp. Soc. Exptl. Biol. **4**, 92 (1950).
[14] R. W. Sperry, Growth Symposia **10**, 63 (1951).
[15] F. Verzar and P. Weiss, Pflüger's Arch. ges. Physiol. **223**, 671 (1930).
[16] P. Weiss, J. Comp. Neurol. **77**, 131 (1942).
[17] P. Weiss, J. Embryol. Exptl. Morphol. **1**, 181 (1953).
[18] F. C. McLean and M. R. Urist, *Bone* (The University of Chicago Press, Chicago, 1955).
[19] C. Grobstein in *Aspects of Synthesis and Order in Growth*, D. Rudnick, editor (Princeton University Press, Princeton, 1954), p. 233.
[20] I. H. Leopold and F. H. Adler, A.M.A. Arch. Ophthalmol. **37**, 268 (1947).
[21] M. S. McKeehan, J. Exptl. Zool. **117**, 31 (1951).
[22] P. Weiss, Quart. Rev. Biol. **25**, 177 (1950).
[23] J. Brachet, *Chemical Embryology* (Interscience Publishers, Inc., New York, 1950).
[24] P. Weiss, Growth **5**, 163 (1941).
[25] P. Weiss, Yale J. Biol. Med. **19**, 235 (1947).
[26] A. Tyler, Growth **10**, Suppl., 7 (1947).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

REVIEWS OF MODERN PHYSICS VOLUME 31, NUMBER 2 APRIL, 1959

49
Molecular Organization of the Nerve Fiber*

FRANCIS O. SCHMITT

Department of Biology, Massachusetts Institute of Technology, Cambridge 39, Massachusetts

IN the context of these studies in the biophysical sciences, it is the purpose of this contribution to identify some of the basic problems of nerve that may be effectively studied at the molecular level by the methods of biophysics and biophysical chemistry, and to describe certain nerve structures, such as the myelin sheath, which provide great scope for the application of such methods and which have important application also to such general biological problems as the nature of cell membranes.

The primary function of neurons is the transfer of information from one region to another, the message being propagated along the fiber from cell body to axon terminals by means of a local disturbance or propagated bioelectric impulse. To many, particularly to those studying the activity of the larger nerve units in the processing of information by the central nervous system or in the behavior of the organism as a whole, the individual neuron is but the structural and functional unit in a highly complex network whose basic function involves the all-or-nothing, on-or-off responses of the units. Thus, in neurophysiology, as well as in bioelectrical studies generally, the most elementary problem which could be posed to the molecular biologist or biophysicist would doubtless be the elucidation of the mechanism by which the impulse is propagated along nerve fibers—i.e., the nature of the changes in the properties of the surface film of the axon that underlie excitation and propagation of the action wave.

If this were in fact the only problem to be dealt with, the present contribution would be very short indeed. Little is known about the molecular composition of the axon surface membrane and even less is known about changes that occur in it during propagation of the impulse. However, for several reasons, I do not wish to take such a limited point of view. In the first place, the axon surface membrane does not exist as a separate entity. It is but part of the neuron which, as a complex reacting system, contains many constituents. Even if one were interested only in bioelectric processes, it would be impossible to understand the phenomena going on in the surface membrane without reference to the structural and biochemical properties of the fiber as a whole. Moreover, nerve performs many functions other than that of propagating impulses. Some of these functions are known, some can be guessed, while others are completely unknown.

For example, somewhere in the neuron there are synthesized certain highly active substances, such as acetyl choline and adrenergic substances, which have transmitter functions; i.e., they transmit the impulse across the junction from the nerve fiber to the muscle fiber or other end organ. The site of synthesis of this material in the neuron, its mode of transfer from this site to the nerve endings (possibly enclosed there in vesicles) and its transmission to the innervated tissue are at present very poorly understood. Certain neurons also produce neurosecretions that can be identified cytologically and histochemically and can be demonstrated to pass from the nerve cell body down the axon to the axon terminals. Possibly other hormonal substances are similarly produced and transmitted, but have thus far escaped detection.

Neurons also manifest certain properties, especially during embryogenesis or regeneration, that are highly specific. Thus, each motor fiber grows out and eventually innervates its appropriate muscle. This innervation is not only highly specific with respect to the muscle fiber which is "sought out" and innervated, but also it is necessary for the stability and maintenance of the normal properties of the muscle fiber. Cutting the nerve fiber to a muscle may cause the latter to change its physiological and pharmacological properties markedly, or even to degenerate.

Finally, it is probable that certain aspects of nerve function may involve chiefly the nerve cell or axoplasm without obvious relationship to the axon surface film. Such processes would probably not lend themselves to study or detection by bioelectric methods. For example, as is discussed in more detail below, certain constituents, such as the fibrous protein, seem to be present in all neurons and to represent one of the characteristic features of neurons generally. However, there is still not the foggiest notion as to the biochemical and physiological function of this protein or protein complex.

It is desirable, therefore, that the biophysical and biochemical investigation of the neuron be not limited to bioelectric processes, but be systematic and exhaustive so as to reveal "what is there," particularly at the molecular and macromolecular level. One has only to consider the history of other fundamental biological problems, such as that of muscle contraction, to realize the importance of such a systematic biochemical and biophysical approach. Only when significant advances in knowledge of the composition and molecular organi-

* These studies were aided by a research grant (B-24) from the National Institute of Neurological Diseases and Blindness, of the National Institutes of Health, Public Health Service, U. S. Department of Health, Education, and Welfare; by a contract between the Office of Naval Research, U. S. Department of the Navy, and the Massachusetts Institute of Technology (NR 101-100); and by grants from the Trustees under the wills of Charles A. King and Marjorie King, and from Lou and Genie Marron.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

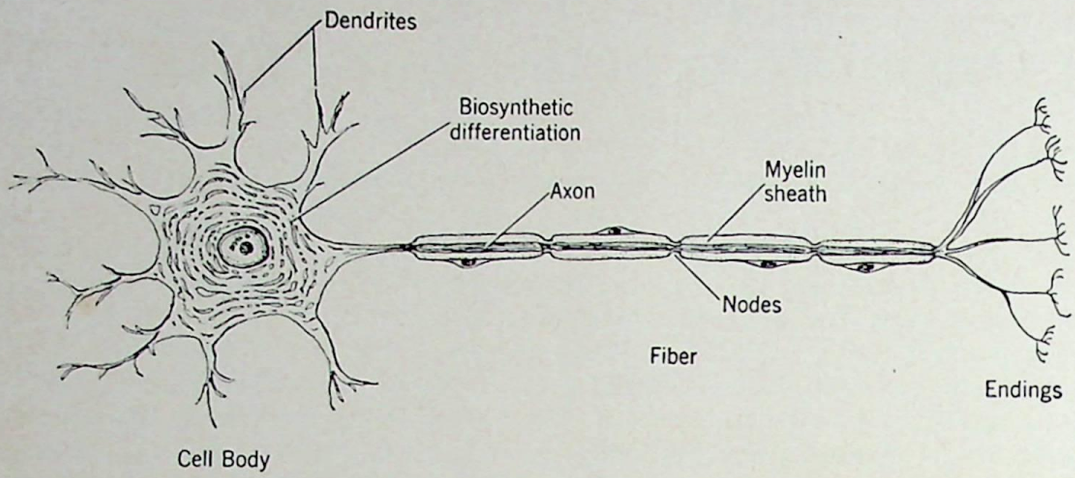


FIG. 1. Diagrammatic representation of a myelinated nerve fiber. Note cell body with dendrites and cytoplasmic organelles, axon, myelin sheath with Schwann cells and nodes of Ranvier, and terminal twigs making synaptic connection with peripheral tissue.

zation of muscle and its fibrous proteins were obtained, could the fundamental problem of the molecular mechanism of contractility be fruitfully attacked. By comparison with muscle, despite much research extending over a century, knowledge of the composition and molecular organization of nerve is still primitive. Moreover, impulse propagation in the nerve fiber is a subtle phenomenon which, unlike contractility in muscle, is not revealed in structural alterations thus far demonstrable even in the electron microscope.

The investigation of the molecular biology of nerve is, therefore, very challenging, particularly to young biophysical scientists not yet steeped in traditional concepts of neurophysiology. This article, which is necessarily limited in space, is meant to be provocative rather than exhaustive in its treatment of basic matter. For more extensive information, the reader is referred to reviews by Schmitt.[1–3]

I. NEURON

The unit of the nervous system, the neuron, consists of a nerve cell body with its dendritic arborization and synaptic connections with the terminal twigs of other neurons; of an axon which is the thin, cylindrical outgrowth of the nerve cell; and of the endings or terminal twigs which make apposition with the innervated cell or tissue through the synapse (see Fig. 1). The neuron has physiological as well as morphological polarization. The impulse passes from presynaptic fiber to nerve cell, down the axon and across the terminal synaptic membrane to the innervated tissue, but not in the reverse direction. For present purposes, it is convenient to consider the cell body, axon, and endings separately.

(A) The Cell Body

The nerve cell body has the task of spinning out the axon in development and of regenerating it in the event of injury. Since the volume of axoplasm to be biosynthesized in a long fiber may be hundreds and, in some cases, thousands of times that of the cell body, it is not unexpected that the metabolism of the nerve cell is very high. To accomplish its impressive task of biosynthesis, the cell is equipped with the cytological differentiation found by electron microscopists characteristically in highly active, secreting gland cells. Particularly prominent are the membrane-limited structures (endoplasmic reticulum, ergastoplasm) richly studded with RNA-rich ribosomes (see Palay[4]). These basophilic structures are identical with the Nissl substance, which has long been studied by cytological methods and by microabsorption spectroscopy and which is known to reflect the synthetic activity of the nerve cell. With improved spectrophotometric methods, such as those developed by Hydén, it has been possible to follow quantitatively the changes in RNA, protein, lipid, and dry weight of regenerating neurons and to correlate these changes with the peripheral growth of the axon (see Brattgård *et al.*[5]). Despite the increase in protein, lipids, and water during the axon outgrowth, the amount of RNA appears to remain constant. However, changes occur in the state of aggregation of RNA; small, finely dispersed particles are formed in the process of "chromatolysis," long known to occur during regeneration. When contact is re-established between the outgrowing axon and the peripheral innervated cell, there is characteristically an increase of RNA and of cell volume until the original values are restored. According to these authors, the only type of cell that can compete with the nerve cell in biosynthetic activity and with the degree of cytological differentiation for such processes, particularly that involving the RNA system, is the exocrine cell of the pancreas of small animals. The neuron is visualized as "an enormous gland cell structure whose lively protein metabolism serves the specific nerve function."[5]

The high metabolic activity of nerve cells (as distinguished from nerve fibers), long known to be characteristic also of normal, unregenerating cells (as in the brain), may be owing in part to the constant synthesis of axoplasm which—according to Weiss and Hiscoe[6] and supported recently by Waelsch[7] and by Ochs and Burger[8]—passes continuously, during the life of the neuron, as a plastic, gelled mass down the axon at a slow rate (*ca* 1 mm/day). The manner in which axoplasm disappears peripherally, e.g., by oxidation or

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

other metabolic process, is not known. If nerve cells are in fact constantly synthesizing axoplasm, it becomes important to establish the purpose served by such a metabolically expensive process. It is possible that one such purpose may be the maintenance of the turgor of the axon over its long extent to the endings. One constituent that would also be constantly synthesized in such a process is the fibrous protein of the axon. When the function of this protein is discovered, light may be shed also on the apparent necessity for its constant synthesis.

The metabolism and the cytological differentiation of the nerve cell are responsive not only to the need for active biosynthesis, as in the regeneration of a severed axon, but also to physiological activity of the neuron. This apparent energy dissipation is in contrast to the highly conservative nature of the process of impulse propagation in the peripheral fiber.

The evidence for the highly developed membrane-limited structures in the cytoplasm of living nerve cells and for the presence of fibrous protein oriented with long axes parallel to the distinguishing directions of the nerve cell is not merely cytological and electron microscopical, but derives also from polarization-optical studies of fresh nerve cells.[9]

(B) Axoplasm and the Axon

As a differentiated type of nerve-cell cytoplasm, axoplasm is a highly complex system. Noteworthy is the low absorption at about 2600 A which characterizes the region peripheral to the axon hillock, roughly at the junction of the nerve-cell cytoplasm with the axon. This is in agreement with the relative scarcity in the axoplasm of ribosomes, as seen in the electron microscope, and of basophilic material, as seen cytologically. Membranous material, possibly consisting chiefly of fragments of the membrane-rich cytoplasmic system, have been observed in squid axoplasm in this laboratory by Maxfield[10] during the process of protein fractionation and physicochemical analysis of axon proteins. Peripheral axoplasm is low in materials absorbing at 2600 A (see von Muralt[11]), consistent with the apparent lack of substantial biosynthesis in this region of the neuron.

Among the particulates commonly present in cytoplasm, only the mitochondria of axoplasm have been subjected to special study. These have properties similar to those isolated from other tissues (Foster[12]). In thin sections of lobster fibers, they have been observed to occur preferentially near the Schwann-cell surface, suggesting the possibility that metabolic energy utilization may occur in this region (Geren and Schmitt[13]).

Although axoplasm is a highly complex, specially differentiated form of nerve-cell cytoplasm, it is highly desirable that the main features of its composition and macromolecular organization be established, not merely for the light it might throw upon the energy coupling involved in active ion transport and in other processes related to impulse propagation, but also as clues to the discovery of other processes carried out by nerves, possibly unrelated to, and not predictable from, those which may be studied by electrical methods.

For such studies of axoplasm, the squid giant fiber is uniquely suited. Axoplasm, relatively uncontaminated by sheath constitutents and other nonaxonic material, can be readily obtained from these fibers. From axoplasm obtained from *Loligo pealii*, the New England squid, analyses have demonstrated (Koéchlin[14]) the presence of previously unknown substances, particularly isethionic acid, which turns out to be the major anion in the acid-base balance of the squid nerve fiber. The fibrous protein of axoplasm has also been studied in this material (Maxfield,[10] Maxfield and Hartley,[15] Schmitt[16]).

More recently, the large squid, *Dosidicus gigas*, that abound in the waters of the Humboldt current off the western coast of South America, have been utilized for chemical investigations. Deffner and Hafter[17] have demonstrated the presence in this axoplasm of some fifteen free amino acids, as well as cysteic amide (a sulfonate closely related to isethionic acid and taurine, which are also present in relative abundance) and other low molecular weight constituents (see Table I), including some five peptides whose composition is now being investigated and whose function is not known. Using high-capacity methods of electrophoretic fractionation and large amounts (10 to 20 g) of freeze-dried axoplasm, these substances are being isolated in quantities sufficient to permit determination of their composition and biological properties.

Among the macromolecular constituents of axoplasm, the fibrous protein is of special interest. Polarization-optical analysis (Bear *et al.*[18]) demonstrated (1) that the fibrous protein exists in the normal fresh axon where it is oriented with long axes parallel to the fiber axis, (2) that the intrinsic birefringence of the protein macromolecules is similar in magnitude to that of other fibrous proteins, and (3) that the partial volume occupied by the oriented fibrous protein is probably relatively small. Electron-microscopic observations[19,20] revealed that the fibrous protein occurs characteristically as thin (*ca* 100 to 200 A) filaments of indefinite length and with no demonstrable axial discontinuities or periodic structure. P. F. Davison and E. W. Taylor in this laboratory have made recent physicochemical studies (unpublished) on fresh axoplasm of *Dosidicus* shipped iced from Chile to Cambridge. From these, it appears that the fibrous protein may have a molecular weight in the order of millions and may be very long and thin. However, the possibility cannot be excluded that the protein as it occurs in the axoplasm so far available (i.e., suspended in cold saline and studied some days after removal from the squid) may in fact be a high polymer, of which the monomer may have a length of the order of 2500 A. In addition, the material which has so far been isolated may represent a complex of two or more proteins. Only when this protein or protein complex has been isolated and characterized chemically and structurally will it be

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

TABLE I. Dialyzable substances in squid axoplasm. Expressed as micromoles per gram dry weight.

Substance	Reaction at physiological pH (~6.7) Acidic Loligo	Acidic Dosidicus	Neutral Loligo	Neutral Dosidicus	Basic Loligo	Basic Dosidicus
Aspartic acid	73.0	74.8				
Glutamic acid	19.6	26.2				
Glycine			10.7	10.4		
Alanine			7.8	9.3		
Serine			3.7	1.0		
Leucine and isoleucine			2.7	0.2		
Valine			2.2	0.5		
Threonine			1.9	0.3		
Proline			1.0	0.1		
Tyrosine			0.7	0.3		
Phenylalanine			0.6	0.15		
Methionine			0.4	0.15		
Arginine					3.2	4.1
Lysine					2.4	0.2
Ornithine					1.8	0.3
Isethionic acid	220.0[a]					
Taurine			98.0	31.0		
Cysteic acid amide			4.4	0.4		
Glycocol, betaine			+	+		
Glycerol			+	+		
Homarine			17.1	18.6		
Peptides			+	+		
Fumaric acid						
Succinic acid	15.0[a]					
K					344.0[a]	+
Na					65.0[a]	+
Ca					7.0[a]	+
Mg					20.0[a]	+
PO_4	16.6[a]	16.1				
SO_4	+	+				
Cl	140.0[a]	+				

[a] Values taken from Koechlin.[14]

possible to determine its biological role. Meanwhile, as it is purified and fractionated, it will be subjected to screening tests intended to provide clues regarding its biological activity.

(C) The Axon Surface Membrane

Electron microscopy permits one to visualize directly the axon surface membrane, though admittedly this is possible only in fibers which have been fixed, embedded, and cut into ultrathin sections. This membrane, or some component of it, presumably corresponds to the irritable membrane whose physical properties before, during, and after the passage of the nerve impulse have been studied by electrophysiologists. Some of the functions that must be subserved by membrane constituents are shown schematically in Fig. 2.

At high resolution, the membrane appears as a double-edged structure with two dark edges separated by a less dense area (see Robertson[21,22]). The total thickness is about 60 to 70 A, the dark and light lines having about equal thickness. The apparent thickness of the dense regions in the membrane may be reduced as methods are improved. A satisfactory interpretation of this structure, in terms of the lipid and protein components of the membrane, must await a detailed study of model systems of lipids and lipid-protein systems. The consensus at present seems to be that the dark regions represent the aqueous, protein-containing interfaces. The less dense regions represent the lipid phase with its preponderant aliphatic hydrocarbon chains. This point of view is not shared by all. Stoeckenius[23,24] believes that the osmium-tetroxide deposits electron-dense material in the region of the unsaturated double bonds of the lipids rather than at the protein, aqueous interface.

Whatever may be the eventual interpretation of the structure of the axon surface membrane, it is perhaps safe to say that the membrane includes a bimolecular

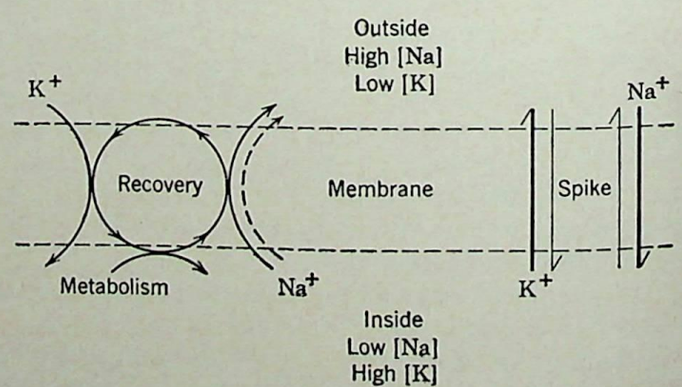


FIG. 2. Diagrammatic illustration of ion movements through the nerve membrane [after A. L. Hodgkin and R. D. Keynes, J. Physiol. 128, 28 (1955)]. Ion movements occurring during the action wave are shown at the right; active transport, involving coupling of metabolic energy, is shown on the left. The excitable membrane pictured is presumably identical with the limiting surface membrane of the axon, or some portion thereof.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

leaflet of mixed lipids and at least a monolayer of protein or other hydrophilic, osmiophilic material on either side of the lipid layer. This represents the "unit" membrane structure, thought by many competent electron microscopists to be characteristic of biological membranes generally. In addition to this structure, which might be thought of as the "floor space" of the cellular factory, one supposes that there are mounted upon or within this membrane the enzymes responsible for the special molecular and biochemical processes which underlie the rapid change in cation permeability at the time of impulse propagation and the "pumping" of sodium ions from axoplasm against an activity gradient into extracellular space. Obviously, much research is required before the details of the molecular organization of the membrane can be established. Meanwhile, in describing the membrane structure, it is perhaps desirable to avoid terms such as "triple" membrane (so-called because of the two dense and one less dense regions); this might cause investigators of the nature of the true "physiological" membrane (i.e., that deduced from indirect, chiefly electrical, methods) to identify one or another component of the fixation artifact seen in the electron microscope as a self-sufficient entity, ignoring the fact that the membrane must almost certainly be considered as a complex system of interacting, interdependent components. Precisely in such situations are the methods of molecular biology different from those of descriptive morphology.

(D) Nerve Endings and the Synapse

At axon endings, electron microscopy reveals an accumulation of particulates, particularly mitochondria. Also characteristic is the presence of small (*ca* 200 to 500 A) vesicles, appearing sometimes in large quantity. It has been suggested that these vesicles contain acetyl choline, adrenalin-like materials, and other physiologically and pharmacologically active substances such as histamine and serotonin. If transmitter substances occur in packets and if such packets were presented at the end organ, perhaps as a secretion across the membrane, certain electrophysiological studies of synaptic potentials might thus find explanation (Katz, p. 524). At present, the origin of the vesicles is still unknown, as are the details of the mechanism by which they or their contents find their way into the endings.

Of considerable theoretical significance is the problem of the nature of the relationship between the limiting membranes of the nerve terminals and of the muscle or other innervated cell. If, as current consensus indicates, there is a space of at least a few hundred Ångström units between these membranes, the problem of the mechanism whereby transmitter substances pass from neuron to innervated tissue is a different one than if the two membranes were strictly in molecular contact. This is still a matter for discussion. The factors involved in determining membrane separation by electron microscopy are considered below in more detail.

II. SATELLITE-CELL DIFFERENTIATIONS AND RELATIONSHIP TO THE AXON

Extending to the farthest reaches of the body, portions of the nerve axon may be very remote from the metabolic center near the nucleus of the cell body. This would constitute an inherently very unstable system. However, the axon is probably seldom actually remote from metabolic centers because it is everywhere surrounded by thin satellite cells, the Schwann cells, which are themselves metabolically active. These cells may well have intimate biochemical relationship with the axon. Such cell-to-cell biochemical relationship and interdependence have in fact been demonstrated *in vitro*. Puck,[25] for example, has shown that a monolayer of cells of a certain type, which are themselves unable to obtain all of their nutritional requirements from the medium, can be stabilized by depositing upon them a layer of cells which can provide or utilize the required factor. These act as "feeder" cells to the underlying cells and constitute a stable system. It seems possible that the evolutionary advance to the higher organisms, with their extensive nervous system, depended upon the development of the process whereby satellite cells come to envelop outgrowing axons, thus permitting the fibers to extend to the extremities of the body as stable neurons.

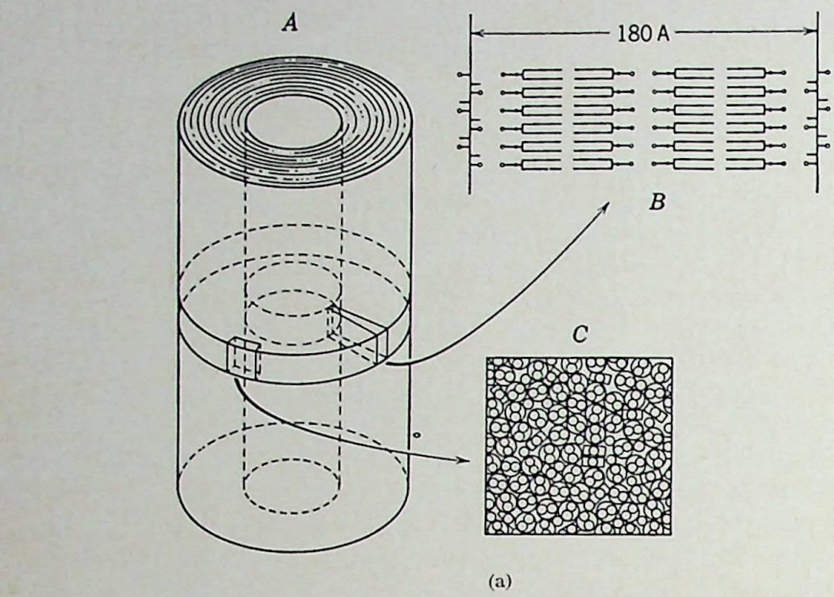
The satellite cells may function in relation to the axon in other ways also. In any homogeneous group of nerve fibers, the velocity of impulse propagation is a direct function of the diameter of the fiber. In invertebrate fibers, to achieve relatively high velocity of propagation (30 to 40 m/sec), the axon diameter becomes relatively enormous (0.5 to 1.0 mm in the squid). However, in myelinated fibers in which the axon is covered to some 99% by myelin, impulse-propagation velocity may be high (50 to 100 m/sec) in fibers of modest diameter (10 to 20 μ). In these cases, impulse propagation is believed to be saltatory in nature rather than continuous; i.e., the membrane changes responsible for excitation occur in very restricted regions (at the nodes of Ranvier), and internodal segments respond as entire units. The investigation of the molecular organization of the myelin, which acts as internodal insulation, and of its mode of formation by the Schwann cell, present a most rewarding opportunity for the methods in the molecular biology of nerve.

(A) Myelin

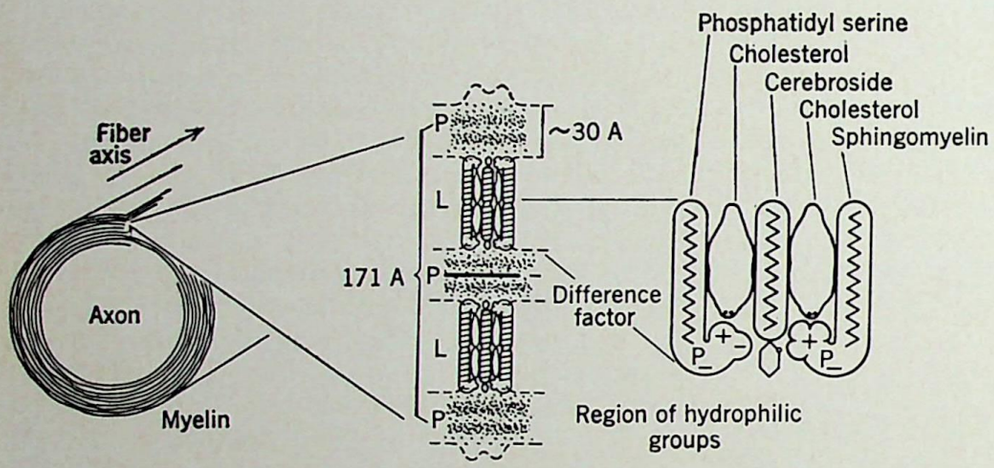
From polarization-optical analyses, it was shown by Schmidt[26] that the lipid molecules in the myelin sheath are oriented with their paraffin chains extending radially, whereas the protein components are oriented with their distinguishing directions perpendicular to this direction, i.e., tangential. The lipid-protein layers must be thin with respect to the wavelength of light, constituting a Wiener mixed body.[27]

X-ray diffraction studies supported this view and

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA



(a)



(b)

FIG. 3. (a) Schematic representation of molecular organization of the nerve myelin sheath (after Schmitt, Bear, and Palmer[28]). *A*, concentric lipid-protein layered structure deduced from polarization optics and from small-angle x-ray diffraction; *B*, unit lipid-protein structure in the radial direction (perpendicular to the planes of the layers); *C*, sections through the molecular chains in the planes of the layers. (b) Diagram summarizing molecular organization of nerve myelin from x-ray diffraction data, as suggested by Finean.[34]

small-angle diffraction showed that the lipid-protein layers have a characteristic thickness of about 171 A in cold-blooded animals and about 185 to 190 A in warm-blooded animals (Schmitt *et al.*[28]). On the basis of extensive studies of the individual lipid types both dry and in aqueous systems, these authors showed that the identity period in the radial direction must include two bimolecular leaflets of mixed lipids, and that the protein and water components must occupy approximately 25 A each. Several alternative models involving bimolecular leaflets flanked by monolayers of protein were proposed (see Fig. 3). It was pointed out[29] that nerve myelin, being susceptible to x-ray diffraction analysis and having a lipid-protein architecture similar to that of cell membranes, might provide a model for the study of such membranes which, because of their thin paucimolecular nature, could not be similarly studied.

The concentric-layered structure of myelin was confirmed by direct visualization in the electron microscope (see Fernández-Morán[30] and Sjöstrand[31]) after development of the technique of thin sectioning for electron microscopy. The thickness of the layers was somewhat less than that demonstrated by x-ray diffraction for the fresh fibers. This difference was shown to be owing to the fixation, dehydration, and embedding required in the electron-optical technique; nerves were subjected to diffraction analysis after each step in the preparative technique, and the thickness of the layers was determined by electron-microscopic examination of thin sections of the same specimens.[32–34]

After the main features of myelin architecture had been determined, the question arose as to how such a structure is produced by the Schwann cell, in collaboration with the axon. A clear-cut answer to this question was provided by Geren,[35,36] at least in the case of peripheral fibers. From observations of embryonic sciatic nerves in chick embryos,[37] Geren proposed her membrane theory which is schematically illustrated in Fig. 4. According to this ingenious theory, now verified by the observations of many investigators (e.g., Robertson[38]), the Schwann cells, after enveloping the outgrowing axons and causing their respective surface membranes to adhere firmly, wrap themselves many times around the axon by a continuous infolding of the outer Schwann-cell surface membrane. The mechanism of myelination in the central nervous system is less clear. Although

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

various authors have seen the myelin layers as a spiral (in transverse section), as in peripheral fibers, others (Luse[39] and DeRobertis *et al.*[40]) have obtained evidence which they interpret as indicating that lipid-protein layers are formed in the cytoplasm of the oligodendrocyte, perhaps by the fusion of vesicles 200 to 800 A in diameter, which condense into the compact myelin layers. The myelin layers are hence not concentric but are spirally wrapped about the axon. After many wrappings (or few, depending upon the fiber type) the Schwann-cell cytoplasm is pressed from among the layers which become condensed into the smectic, paracrystalline material of the finished myelin. This process requires a prolific biosynthesis of lipid-protein material. Where this is synthesized in the cell and how it comes to be incorporated in the cell membrane is not yet known. The myelination process would seem to provide an ideal system in which to study the molecular mechanism of lipid-protein synthesis and of membrane formation in general.

From Geren's membrane theory, it immediately becomes obvious why there are *two*, rather than one, bimolecular layers of lipids in the x-ray identity period in the radial direction (see Schmitt, Bear, and Palmer[28]). As the outer Schwann-cell membrane infolds, the outer surfaces of the membrane remain in contact. As the layers condense, with the expulsion of Schwann-cell cytoplasm, the membrane surfaces which faced the Schwann-cell cytoplasm also adhere, fuse, and form the darkly staining line seen in the electron micrographs (see Fig. 5). The alternate less-dense line represents the fusion of membrane surfaces which originally faced outward to extracellular space. Finean,[34] who has continued the x-ray analysis in a systematic fashion over the last decade, has proposed a molecular model of myelin in which the cholesterol molecules fit in between the phospholipid and cerebroside molecules in characteristic fashion (see Fig. 3). Further refinements in the x-ray and electron-microscopic techniques, together with the study of various lipid-protein models, may provide important new information about membrane structure which will be of vital importance in the rapidly expanding field of molecular biology.

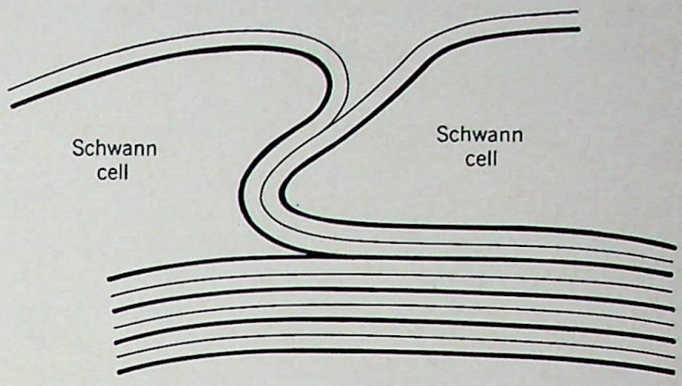


FIG. 5. Illustrating enfolding of Schwann-cell surface membranes to form compact myelin, after the theory of Geren.[35] Note differentiation between inner and outer surfaces of enfolding membrane; in the double membrane, this gives rise to Finean's "difference factor" and explains why the full repeat distance in the radial direction involves two bimolecular leaflets of lipo-protein (two membranes).

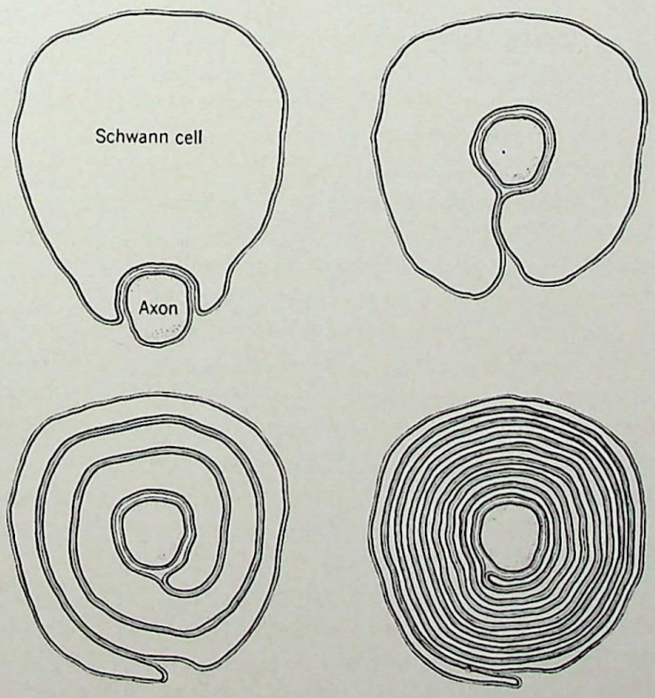


FIG. 4. Schematic representation of the membrane theory of the origin of nerve myelin, (after Geren[35]). Four stages in the engulfment of the axon by the Schwann cell and the wrapping of many double layers which, after condensation, form the compact myelin.

As Schwann or satellite cells envelop the outgrowing axon and start to wrap their surfaces in multiple folds about it, adjacent Schwann cells come into contact and form the nodes of Ranvier. This nodal region of the finished, myelinated fiber has not yet been exhaustively studied in the electron microscope. In view of the critical nature of this region, suggested by the theory of saltatory conduction, its ultrastructure merits careful study. According to a preliminary report by Robertson,[41] the Schwann cells on either side of the node form finger-like processes which interdigitate closely over the node. The region between the processes, presumably containing extracellular fluid, is of the order of several hundred Ångström units thick. To this extent, the axon surface membrane is exposed to extracellular fluid at the node, according to Robertson.

The incisures of Schmidt-Lantermann presented a problem because, if they are continuous channels between the axon and the exterior through the myelin, they would offer conducting paths for current. However, electron microscopy has demonstrated that this is not the case.[42,43] The incisures are produced by a systematic shift in the relationship of all of the helically wrapped membranes, but each membrane is continuous across the incisural gap. There are thus as many layers traversing the incisures as occur in the compact region of the myelin,[5] and the incisures are therefore not regions

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

through which ions and currents can freely flow into the axon.

III. SATELLITE-CELL-AXON RELATIONSHIPS IN "UNMYELINATED" FIBERS

It has long been known to neurohistologists, particularly to those of the Spanish school, that nerve fibers are everywhere covered by satellite cells. The fact that these cells and their possible functional role—in the mature as well as in developing and regenerating fibers—have not greatly interested physiologists, or even histologists, is to some extent owing to their small size (*ca* 0.1 to 1 in most invertebrate fibers) or to the fact that myelin (which has been much studied by histologists and physiologists) was not recognized as satellite-cell substance (i.e., a compact helical layering of double surface membranes). The fact that, early in the development of histology, a sharp distinction was made between "myelinated" and "unmyelinated" fibers (primarily on the basis of staining reactions), led some physiologists to the mistaken notion that a substantial fraction of the animals' nerve fibers are in effect naked axons. From this, it is obvious why the cellular covering of the axon attracted little interest except to those studying nerve development or regeneration, where the satellite-cell covering is unquestionably of functional significance.

Polarization-optical studies (see Schmitt and Bear[29]), based largely on a reinterpretation of the mechanism of the so-called metatropic reaction discovered by Göthlin,[44] strongly suggested that at the surface of all nerve fibers, vertebrate or invertebrate, myelinated or unmyelinated, oriented lipid-protein material is present which has qualitative similarity of molecular organization though quite variable quantitatively from one fiber type to the next. This led to the notion that "all nerve fibers are to some extent myelinated."[29]

This matter could be clarified only when the region at the surface of nerve fibers—i.e., the monolayer of satellite cells—could be observed at high resolution by the electron microscope. Details of the discoveries thus brought to light are beyond the scope of this paper, but a few of the more significant aspects at the molecular level, particularly insofar as they bear on basic physiological concepts, are considered.

Investigation with the electron microscope of the detailed structure of the satellite cells and their relation to the axon is not yet a decade old, and it is, therefore, not unexpected that knowledge of the subject is still very primitive. No doubt, special methods of fixation and preparation will have to be developed to permit investigation of the most fundamental aspects of the problem, such as the possible variation of intercellular relationship, as manifested morphologically by the apparent spaces or channels between axon surface and satellite cells and between satellite cells themselves, with function.

The typical invertebrate, unmyelinated (but not unsatellited!) nerve fibers, such as those of the lobster leg or claw nerve, are surrounded by satellite cells which are very thin (*ca* 0.1 μ). Except at the region of the nuclei, there is relatively little space between the two parallel surfaces of the cells (i.e., facing the axon surface membrane and that facing the basement membrane, connective tissue investment), although mitochondria and some of the typical equipment of the cytoplasm of active cells may be seen. The axon surface is characteristically highly contorted, and axoplasmic mitochondria are seen with the long axes of their ellipsoids oriented into the folds of these contortions, suggestive of energy utilization at this interface (see Fig. 6). The outpocketings of the satellite cells into the axon may manifest all of the stages to be expected if particulates (mitochondria?) were actually being formed and extruded into the axon by filling such outpocketing with satellite-cell cytoplasm (including its particulates) and pinching them off into the axoplasm (see Geren and Schmitt[13,45]).

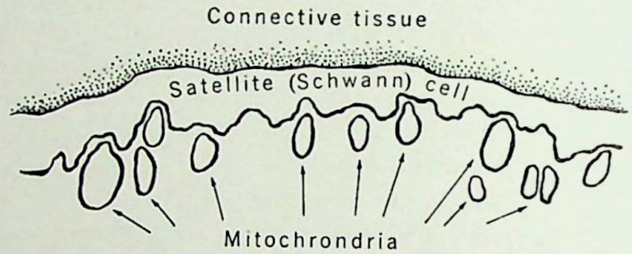


FIG. 6. Diagrammatic illustration of thinness of satellite cell in lobster fiber and of intimate relationship between mitochondria and axon surface membrane (after Geren and Schmitt[41]).

Such very thin cells would be expected to show only a very weak birefringence and metatropic effect, since only two lipid-protein surface membranes are involved. Nevertheless, this optical effect was observed and correctly interpreted, except that the layers were simply assigned to the region at the surface of the axon, rather than to the membranes of the satellite cells.[46]

The satellite cells of squid giant fibers are somewhat thicker (*ca* 1 μ) than those of lobster, though the ratio of thickness to axon diameter may actually be less. In any histological section of the giant fibers, as many as two or three nuclei may be present. If one were to chart the boundaries and domains of the satellite cells upon the axon surface membrane (e.g., by slitting the fiber and flattening out the membrane), one would see that many cells are involved in the formation of the cellular monolayer that surrounds the axon.

The most distinguishing feature of the cytoplasm of these cells is the presence of many (a half-dozen or more) osmiophilic, dense membranous layers lying in planes predominantly parallel with the cell surface. Whether these layers are related to those present in typical, metabolically active cells ("endoplasmic reticulum"), or whether they have a special structure and function in satellite cells remains to be established by more-detailed examination. However, if one assumes that the layers have the composition and ultrastructure typical

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

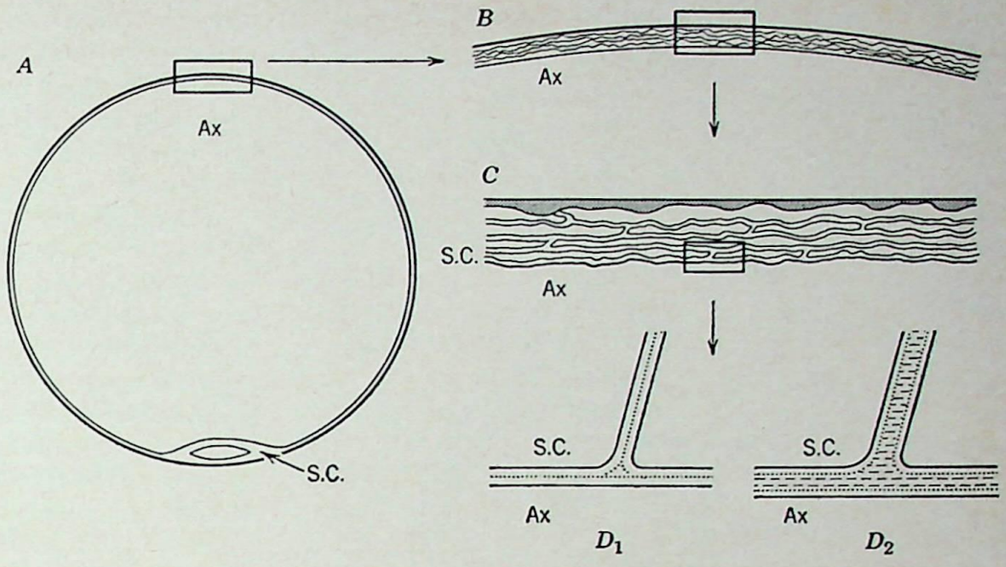


FIG. 7. Relationship of axon (Ax) and Schwann cell (S.C.) in the squid giant fiber (after Schmitt and Geschwind[2]). *A*, cross section of giant fiber showing very thin Schwann cell surrounding axon. ×125. *B*, enlarged segment portion of Schwann cell, ×3000; *C*, same as *B*, showing intracytoplasmic layers, ×9000; and D_1 and D_2, possible relationships at the molecular level of axon surface membrane and Schwann-cell membrane. D_1 presumes layers are in molecular contact. (Water layer does not exceed about 15 A in thickness.) D_2, showing thicker water layers (100 to 200 A) between adjacent membranes. ×150 000.

of lipid-protein membranes generally (hence, also of myelin), it is possible to account for the polarization-optical properties described by Bear *et al.*[47] and ascribed by them to a metatropic layer at the outer surface of the satellite cells—a kind of myelin-like layer. Geren and Schmitt[45] have shown that the optical effects can satisfactorily be accounted for on the basis of the intracytoplasmic dense layers present in the satellite cells. There is no differentiated myelin-like layer at the outside of these cells, only the amorphous basement membrane.

One of the structural details that must await further investigation for clarification, but that may be of great ultrastructural and perhaps functional significance, is the nature of the boundaries between satellite cells and the axon surface membrane, between adjacent satellite cells and between intracytoplasmic membranes. Of particular importance is the thickness of the aqueous regions between these membranes, because ions passing from extracellular space to axoplasm and in the reverse direction must traverse such channels.

The situation is illustrated schematically in Figs. 7 and 8. If the adhesion between the surface membranes were as close as in the myelin sheath, there would be no substantial channels for ion diffusion. The aqueous phase would have a thickness of the order of 12 A according to the x-ray diffraction analysis of myelin struc-

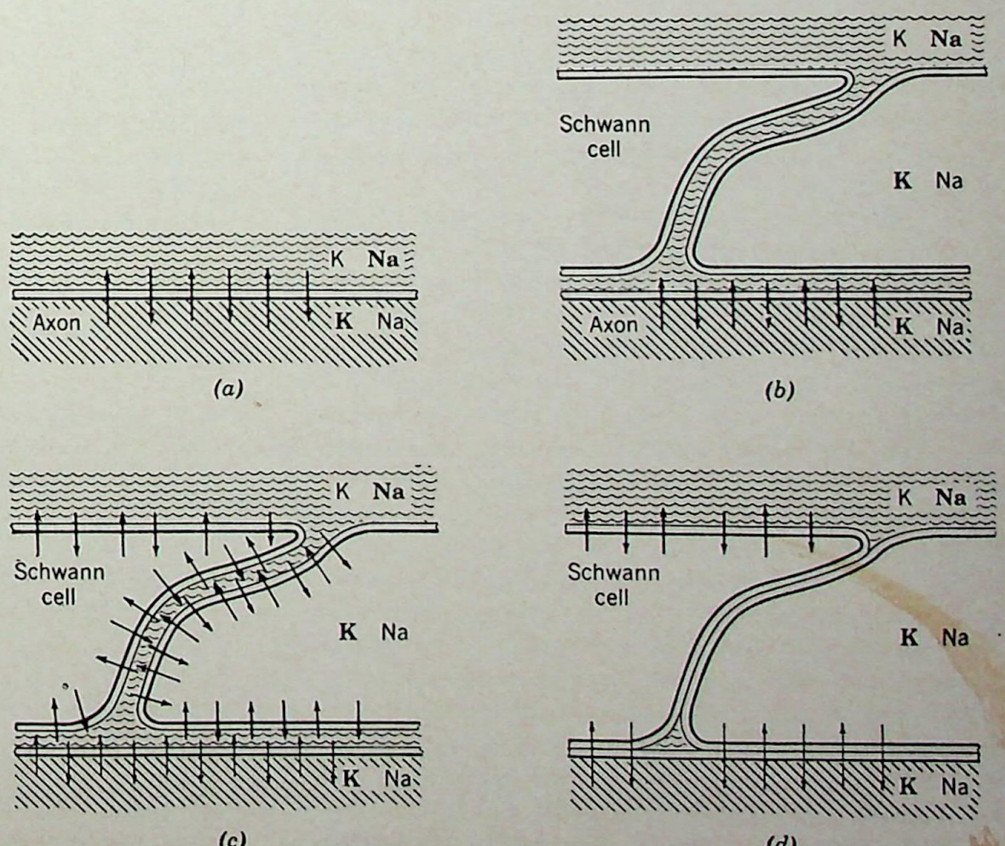


FIG. 8. Possible diffusion pathway for solutes passing between axoplasm and extracellular fluid (after Schmitt[3]). (*a*) bare-axon membrane is exposed to extracellular fluid (*b*) diffusion channel is 150 to 200 A thick, filled with extracellular fluid. Active interchange occurs only between axon surface membrane and fluid in channels; passive diffusion occurs in other channels. (*c*) same as (*b*) except that active interchange of metabolites and solutes occurs across all membrane surfaces; satellite cells participate actively in processes involving interchange between axon and extracellular fluid. (*d*) satellite cells and axon membranes are in molecular contact; water channel not more than 10 to 15 A thick with structure comparable to that of compact myelin.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

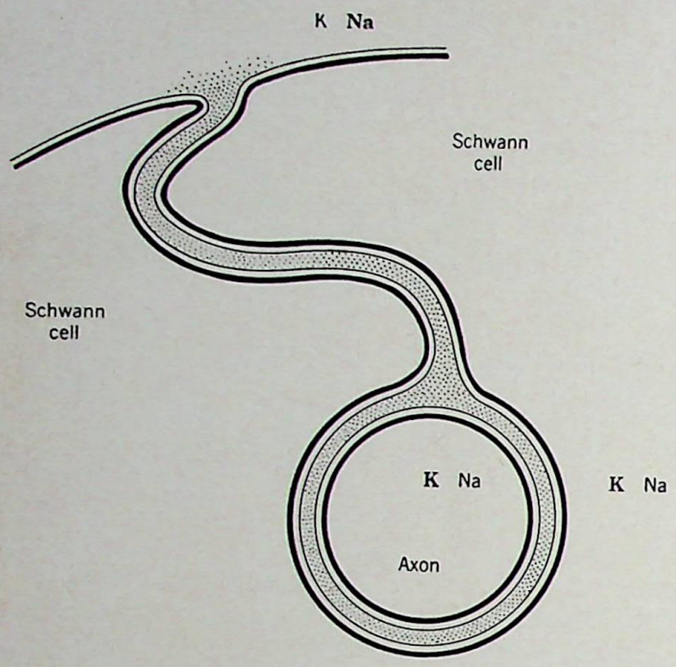


FIG. 9. Diagrammatic representation of relationship of very thin *C* fibers to Schwann cell (after Robertson[48]).

ture.[28] On the other hand, if these intercellular relationships are similar to those found by Robertson[47] and by electron microscopists to obtain between cells generally, the aqueous channels, communicating with the frank extracellular fluid of the ground substance of the surrounding connective tissue, would have a thickness of about 150 A.

Robertson[21,22,48] has found such a membrane relationship in the case of the vertebrate "*C*" fibers. In these nerves, the very thin (*ca* 0.1 μ) axons pass through the satellite (Schwann) cell tube enclosed in a fold of the surface membrane of the satellite cell in a manner analogous to that assumed by Geren's membrane theory to explain the origin of myelin (see also the earlier works of Gasser[49–52]).

In analyzing the functional significance of the intermembrane relationship and the thickness of the aqueous regions between limiting membranes *in life*, several factors must be considered. Indeed, this situation illustrates certain indeterminacies that characterize the electron microscope approach to molecular biology generally.

From the properties of the constituent membranes, including the charge on the membrane molecules, a separation of the order of 100 to 200 A between the membranes might be expected.[53–56] This equilibrium distance will vary with the ionic strength (a statistical physicochemical concept that is difficult to apply to such thin capillary spaces). If the equilibrium distance between surface membranes is responsive to the ionic strength of the environment, it is obvious that a factor of importance is the ionic strength of the fluid used to fix the nerves. Another factor is the osmotic pressure normally characterizing the intercellular fluid. If this is not identical with that of the intracellular fluid, then when the system is placed in the fixative (which presumably quickly destroys membrane semipermeability and such vital equipment as "ion pumps"), an osmotic transfer of water may be expected, and this might importantly influence the distance between surface membranes as observed in the electron microscope after fixation and sectioning.

It has become evident from electron-microscopical studies of metabolically active cells not only that the metabolic activity involves the presence of membranes separated by small distances, but also that the products of the metabolism, presumably including ions, must find their way into the aqueous space between the membranes. The presence of these metabolites in turn must affect the equilibrium separation between the membranes and the properties of the intercellular fluid within these thin capillary spaces. Such factors must be taken into consideration in dealing realistically with these spaces which are the channels of communication between the axon surface membrane (presumably the seat of the rapid alterations responsible for the propagation of the impulse in the nerve fibers) and frank extracellular space. According to Frankenhaeuser and Hodgkin,[57] if such intercellular channels are the route of diffusion of potassium ions from axoplasm to the exterior, the electrical data would require a channel thickness of about 300 A in a structure similar to that shown by Geren and Schmitt[13] assuming a process of passive diffusion calculable on the basis of properties characteristic of macroscopic bulk phases. Whether or not additional properties, deriving from the microscopic dimensions of the intermembrane system and from the presence of solutes, particularly metabolic products of the satellite cells, require modification of this view remains for further analysis to determine. That the satellite cells of the squid giant fiber are indeed metabolically active has been shown by the measurements of oxygen consumption of the sheath cells after removal of the axoplasm,[58] and from an assay of enzyme activities of sheath cells and axoplasm.[59]

BIBLIOGRAPHY

[1] F. O. Schmitt in *Metabolism of the Nervous System*, D. Richter, editor (Pergamon Press, London, 1957), p. 35.

[2] F. O. Schmitt and N. Geschwind, Progr. in Biophys. and Biophys. Chem. **8**, 166 (1957).

[3] F. O. Schmitt, Exptl. Cell Research **5**, Suppl., 33 (1958).

[4] S. L. Palay in *Ultrastructure and Cellular Chemistry of Neural Tissue*, H. Waelsch, editor (Paul B. Hoeber, Inc., and Harper and Brothers, New York, 1957), p. 31.

[5] S. O. Brattgård, J. E. Edström, and H. Hydén, Exptl. Cell Research **5**, Suppl., 185 (1958).

[6] P. A. Weiss and H. P. Hiscoe, J. Exptl. Zool. **107**, 315 (1948).

[7] H. Waelsch, J. Nervous and Mental Diseases **126**, 33 (1958).

[8] S. Ochs and E. Burger, Am. J. Physiol. **194**, 499 (1958).

[9] P. Chinn, J. Cellular Comp. Physiol. **12**, 1 (1938).

[10] M. Maxfield, J. Gen. Physiol. **37**, 201 (1953).

[11] A. von Muralt, *Die Signalübermittlung im Nerven* (Birkhäuser, Basel, 1946).

[12] J. M. Foster, J. Neurochem. **1**, 84 (1956).

[13] B. B. Geren and F. O. Schmitt, Proc. Natl. Acad. Sci. U. S. **40**, 863 (1954).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

[14] B. A. Koéchlin, J. Biophys. Biochem. Cytol. **1**, 511 (1955).
[15] M. Maxfield and R. W. Hartley, Biochim. et Biophys. Acta **24**, 83 (1957).
[16] F. O. Schmitt, J. Cellular Comp. Physiol. **49**, Suppl. 1, 165 (1957).
[17] G. Deffner and R. Hafter, Biochim. et Biophys. Acta (to be published).
[18] R. S. Bear, F. O. Schmitt, and J. Z. Young, Proc. Roy. Soc. (London) **B123**, 505 (1937).
[19] F. O. Schmitt, J. Exptl. Zool. **113**, 499 (1950).
[20] F. O. Schmitt and B. B. Geren, J. Exptl. Med. **91**, 499 (1950).
[21] J. D. Robertson in *Ultrastructure and Cellular Chemistry of Neural Tissue*, H. Waelsch, editor (Paul B. Hoeber, Inc., and Harper and Brothers, New York, 1957), p. 1.
[22] J. D. Robertson, J. Biophys. Biochem. Cytol. **4**, 349 (1958).
[23] W. Stoeckenius, Exptl. Cell Research. **13**, 410 (1957).
[24] W. Stoeckenius in *Proceedings of the Fourth International Congress on Electron Microscopy* (Springer-Verlag, Berlin, to be published).
[25] H. W. Fisher and T. T. Puck, Proc. Natl. Acad. Sci. U. S. **42**, 900 (1956).
[26] W. J. Schmidt, *Die Doppelbrechung von Karyoplasma, Zytoplasma und Metaplasma* (Gebrüder Borntraeger, Berlin, 1937).
[27] O. Wiener, Abhandl. sächs. Ges. Wiss. math. phys. Kl. **32**, 509 (1912).
[28] F. O. Schmitt, R. S. Bear, and K. J. Palmer, J. Cellular Comp. Physiol. **18**, 31 (1941).
[29] F. O. Schmitt and R. S. Bear, Biol. Revs. **14**, 27 (1939).
[30] H. Fernández-Morán, Exptl. Cell Research **1**, 309 (1950).
[31] F. S. Sjöstrand, Experientia **9**, 68 (1953).
[32] J. B. Finean, F. S. Sjöstrand, and E. Steinmann, Exptl. Cell Research **5**, 557 (1953).
[33] H. Fernández-Morán and J. B. Finean, J. Biophys. Biochem. Cytol. **3**, 119 (1957).
[34] J. B. Finean, Exptl. Cell Research **5**, Suppl., 18 (1958).
[35] B. B. Geren, Exptl. Cell Research **7**, 558 (1954).
[36] B. B. Geren in *Cellular Mechanisms in Differentiation and Growth*, D. Rudnick, editor (Princeton University Press, Princeton, 1956), p. 213.
[37] B. B. Geren and J. Raskind, Proc. Natl. Acad. Sci. U. S. **39**, 880 (1953).
[38] J. D. Robertson, J. Biophys. Biochem. Cytol. **1**, 271 (1955).
[39] S. A. Luse, J. Biophys. Biochem. Cytol. **2**, 777 (1956).
[40] E. DeRobertis, H. M. Gerschenfeld, and F. Wald, J. Biophys. Biochem. Cytol. **4**, 651 (1958).
[41] J. D. Robertson, J. Physiol. **135**, 56*P* (1957).
[42] M. Luxoro, Proc. Natl. Acad. Sci. U. S. **44**, 152 (1958).
[43] J. D. Robertson, J. Biophys. Biochem. Cytol. **4**, 39 (1958).
[44] G. F. Göthlin, Kgl. Svenska Vetenskapsakad. Handl. **51**, 1 (1913).
[45] B. B. Geren and F. O. Schmitt in *Symposium on Fine Structure of Cells* (Interscience Publishers, Inc., New York, 1955), p. 251.
[46] R. S. Bear and F. O. Schmitt, J. Cellular Comp. Physiol. **9**, 275 (1937).
[47] R. S. Bear, F. O. Schmitt, and J. Z. Young, Proc. Roy. Soc. (London) **B123**, 496 (1937).
[48] J. D. Robertson, J. Biophys. Biochem. Cytol. **3**, 1043 (1957).
[49] H. S. Gasser, Cold Spring Harbor Symposia Quant. Biol. **17**, 32 (1952).
[50] H. S. Gasser, J. Gen. Physiol. **38**, 709 (1955).
[51] *Ibid.*, **39**, 473 (1956).
[52] H. S. Gasser, Exptl. Cell Research **5**, Suppl., 3 (1958).
[53] K. J. Palmer and F. O. Schmitt, J. Cellular Comp. Physiol. **17**, 385 (1941).
[54] K. Hess and H. Kiessig, Chem. Ber. **81**, 327 (1948).
[55] H. L. Booij and D. Bungenberg-de Jong in *Protoplasmatologia Handbuch der Protoplasmaforschung* (Springer-Verlag, Vienna, 1956).
[56] J. Th. G. Overbeek, unpublished lecture, "Physical chemical properties of layered systems," given before the Electron Microscope Society of America (10 September 1957).
[57] B. Frankenhaeuser and A. L. Hodgkin, J. Physiol. **131**, 341 (1956).
[58] R. R. Coelho, J. W. Goodman, and M. B. Bowers (manuscript in preparation).
[59] N. R. Roberts, O. Lowry, E. J. Crawford, and R. Coelho, J. Neurochem. **3**, 109 (1958).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

50
Nature of the Nerve Impulse

BERNARD KATZ

Department of Biophysics, University College, London, W. C. 1, England

INTRODUCTION

THE mechanism by which signals are rapidly transmitted over long distances in the body has formed one of the principal preoccupations of biophysicists for well over a century. Since the days of Du Bois-Reymond (about 1850),[1] it was known that nerves not only are excited by electric currents but also produce, in the course of their activity, electric currents of their own. His pupil, Hermann,[2] suggested that the propagation of a signal along a nerve fiber (or axon) involves a process of recurrent electric stimulation from point to point. Hermann also called attention to the apparent cable properties of nerve fibers, although he recognized that, by passive cable transmission alone, nerve signals could not travel over any appreciable length. Later, under the influence of Nernst, Bernstein[3] developed the basis of the modern membrane theory. According to Bernstein, the key to our understanding of the nerve impulse is to be found in the properties of the surface membrane of the axon, and especially its selective and changeable permeability (or conductance) to ambient ions. His theory has been modified in important details, but in essence it still stands, and the advances which have been made in this field during the last two decades have confirmed the usefulness of Bernstein's basic concept.

The view which now has emerged is (a) that the electrical events in a nerve fiber are governed by the differential permeability of its surface membrane to sodium and potassium ions, and (b) that these permeabilities themselves depend upon the electric field across the surface. The interaction of these two factors leads at a certain critical threshold level to excitation, that is, to a regenerative release of electrical energy from the axon membrane, and the propagation of this change along the fiber in the form of a brief, all-or-none, electrical impulse (the so-called spike or action potential).

As F. O. Schmitt has pointed out (p. 455), when a physiologist speaks of the axon membrane, he usually has in mind an abstraction rather than a microscopically identified structure. This point, *viz.*, the relation between fine structure and functional properties, is discussed later in this paper; here the main concern is to indicate what evidence there is for such physiological "abstractions."

RESTING NERVE FIBER

For the purpose of discussing the principal features of the nerve signal, the structural picture can be reduced to that of a long cylindrical tube with a surface membrane which separates two aqueous solutions of equal osmolarity but of different chemical composition. In the external medium, more than 90% of the osmotic balance is made up of sodium and chloride ions; while inside the cell, these ions account for less than 10% of the solutes, potassium taking the place of sodium, and various organic anions (which presumably are synthetized within the cell itself) taking the place of chloride. Our main concern is with the concentration differences of sodium (about 10 times higher outside) and potassium (about 30 times higher inside) across the cell surface. With the use of very fine KCl-filled micropipettes, it has been possible to penetrate the fiber surface without serious damage and to measure the electric PD between inside and outside. There is a potential difference across the membrane of some 60 to 90 mv (inside negative), while there are no detectable potential differences within the interior of the normal resting cell.

Observations of this kind have been made on a variety of nerve and muscle fibers. Most of the evidence described in the following has been obtained from the giant axon of the squid whose large size (nearly 1 mm thick) allows one to introduce an assembly of fine electrodes along the inside to analyze the ionic content of a single fiber and to observe the movement of radio-tracers within selected regions of such fibers (see Hodgkin,[4,5] also Huxley[6] and Katz[7] for review and references).

It should be mentioned perhaps that the nerve fibers of squid and other cold-blooded animals can be isolated and kept in a suitable salt solution. They remain in a functioning condition for many hours during which they can be made to propagate many thousands of impulses of approximately the same voltage and velocity as initially, *in situ* and with the blood circulation intact. Such fibers remain capable, to some extent, of replenishing the losses from their chemical stores which are drawn upon during periods of activity. Nevertheless, it is true that isolated tissues and, in particular, nerve fibers (which, after all, are only peripheral stumps of a cell) are no longer in a steady state, and their electrochemical accumulator gradually runs down.

It is of interest that muscle fibers can be used to study propagation of impulses as well as nerve. Muscle fibers, unlike severed nerve axons, are self-contained cell units which possess a built-in apparatus of distributed nuclei (and which, incidentally, do not possess investing layers of satellite cells). The surface membrane of a muscle fiber has certain important properties in common with that of a nerve: it provides a selective barrier between cytoplasm and surroundings, and it serves to conduct electric excitation rapidly over the whole length from

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

the nerve-muscle junction to the tendon. This is an essential mechanism for the skeletal muscle, for the action potential provides the local stimulus to the contractile process at each point of the fiber, and a high speed of propagation is needed to elicit an efficient synchronous twitch. As far as has been ascertained, the properties of the vertebrate-muscle impulse are exactly analogous to those of the nerve action potential.

MAINTENANCE OF THE STEADY STATE

The first question to consider is: What are the mechanisms responsible for the maintenance of a steady state in a resting (i.e., nonstimulated) nerve or muscle cell? In particular, how are the electrical and ionic concentration differences kept up between interior and surroundings?

Is potassium chemically bound? It has been suggested from time to time that one does not require selective membrane properties in order to explain the preferential accumulation of potassium in the cytoplasm. It would be sufficient, for example, if the cellular proteins had a special chemical affinity to potassium rather than to sodium. This idea of chemical binding of potassium seemed unlikely on several grounds: it would be difficult to account for the osmotic pressure and electric conductivity of the cell contents unless at least a large proportion of the intracellular potassium was present in the form of free ions. The most direct evidence was obtained by Hodgkin and Keynes[8] in the following experiment. A region of an isolated *Sepia* axon (about 150 μ thick) was bathed in a droplet containing radiopotassium. After a quantity of the labeled ions had entered the axoplasm, the external droplet was washed away. It may be noted that the mixing between applied tracer and intracellular potassium takes many hours, but the exposure time was sufficient to build up an adequate amount of radioactivity within the axon. Subsequently, the spread of this patch by diffusion along the interior of the fiber was determined; at the same time, a potential difference was applied to the ends of the fiber, and the speed with which the labeled patch moved toward the cathode was measured. The experiment showed that the diffusion coefficient and the electric mobility of the tracer ions were only slightly less inside the cell than in free aqueous solution. There was clearly a highly resistive surface barrier, for the entry of the externally applied potassium into the axoplasm was very slow; but once the tracer ions had passed this barrier, they continued to behave as free ions.

BERNSTEIN MEMBRANE THEORY

The experiment of Hodgkin and Keynes seems to rule out the suggestion of chemical binding of potassium and provides direct evidence for the existence of a resistive surface membrane as envisaged by Bernstein more than 50 years ago. He postulated that the nerve membrane was selectively permeable to potassium alone, and impermeable to sodium, chloride, and the internal anions. This was an attractive hypothesis, for it would explain not only the maintenance of steady ionic concentration gradients, but also the existence of the resting potential as resulting from a potassium-concentration cell. Moreover, it was possible to predict an inverse logarithmic relation between the external potassium concentration and the membrane PD, and over a fairly wide range there was satisfactory agreement between theory and experiment. Bernstein went one step further. He postulated that the membrane permeability to the other ions increased during electric excitation so that the impulse should be accompanied by a sharp increase of the membrane conductance and a drop toward zero of the membrane potential (or rather to the low level of a liquid junction PD between axoplasm and external fluid). For many years, these ideas were accepted as a reasonable explanation, and the evidence, rather inaccurate and indirect as most of it was, tended to support it. More recently, it was strengthened by the observation of Cole and Curtis[9] that the rising phase of the action potential indeed is accompanied by a rapid 40-fold increase of the ionic membrane conductance.

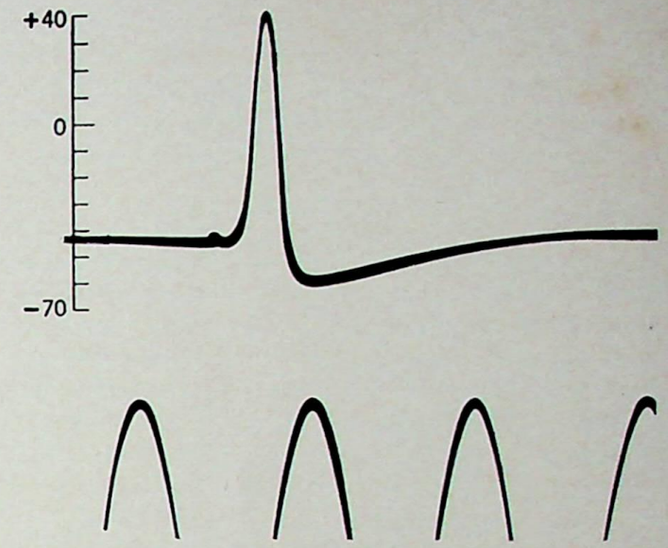


FIG. 1. Action potential recorded with an internal electrode from a squid giant axon. The scale shows the internal potential in millivolts, relative to the outside bath. The time marks are at 2-msec interval [from A. L. Hodgkin and A. F. Huxley, Nature 144, 710 (1939)].

The flaws began to appear soon after, when increasing use was made of the critical methods introduced by Cole and by Hodgkin, *viz.*, the single axon technique and the use of intracellular recording, and also when improved chemical and tracer methods were employed in observing ion movements. In 1939, Hodgkin and Huxley[10] (see also Curtis and Cole[11]) discovered that the membrane potential during the nerve impulse does not move simply toward the zero level (Bernstein's predicted "depolarization"), but reverses substantially to a level of 40 to 50 mv—inside positive (see Fig. 1). This was clearly incompatible with the second part of the Bernstein theory just now outlined. Two years later,

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

Boyle and Conway[12] demonstrated that muscle fibers are permeable to chloride as well as to potassium, and shortly afterward it became clear that even sodium can enter the fibers, though apparently it encounters a much higher membrane resistance than does potassium.

SODIUM PUMP

This raised an altogether new problem, for the demonstration of even a small permeability to *sodium* made it impossible to regard the physical properties of the surface barrier as sufficient to maintain the concentration and potential gradients in a steady state. Potassium and chloride appear to be very nearly in electrochemical equilibrium, their large concentration differences being balanced approximately by the membrane potential. But, in the case of sodium, both driving forces act in the same direction and tend to shift the ion into the interior. If the membrane resistance toward sodium is not infinite, then some mechanism entirely different from the ones discussed so far is required to keep its internal concentration at a steady low level.

It was at this stage that the postulate of a metabolic sodium pump was introduced—that is, some secretory process in which an energy-yielding reaction of the cell is utilized to perform the work of a Maxwell demon, moving sodium ions uphill and outward through the membrane as fast as they leak into the cell in the direction of the electrochemical gradient. There was nothing intrinsically implausible about the suggestion, for the permeability of the resting cell surface to sodium is normally so low that the leakage rate remains very small, and the work required of the pumping process amounts to only a small fraction of the energy which is continuously being made available by the metabolism of the cell.[13]

That there is, in fact, a direct relation between cellular metabolism and sodium extrusion was first shown by Hodgkin and Keynes[14] who observed a cessation of the efflux of tracer sodium when the axon was treated with certain metabolic inhibitors (dinitrophenol, azide, cyanide). This was a reversible change and one which specifically affected the extrusion rate of sodium, not its rate of leakage into the cell. More recently, Caldwell and Keynes[15] have shown that "pumping" is temporarily resumed by the cyanide-inhibited axon if a dose of adenosinetriphosphate or arginine phosphate is injected into the axoplasm.

At one time, it was thought that the sodium pump operated by expelling sodium alone, and in doing so was capable of separating electric charges at the fiber surface, moving them against the existing potential gradient, and so contributing directly to the build-up of the resting emf. This idea has been made unlikely by the findings (a) that the extrusion rate of sodium is unaffected by the potential gradient against which such an electrochemical pump would have to work, and (b) that the stoppage of sodium extrusion by metabolic poisons has no immediate effect on the resting (or action) potential, which runs down very slowly as a result of the gradual diminution of the ionic concentration gradients. It appeared, therefore, that the ion pump works on a principle of electroneutrality, either taking sodium out in company with some anion (e.g., bicarbonate or phosphate), or through a cation-exchange process. There is evidence that the latter is the prevailing mechanism, and that extrusion of sodium and internal uptake of potassium are linked closely in the pumping process. The most suggestive piece of evidence is the finding[14] that the extrusion rate of sodium becomes reduced greatly when potassium is withdrawn from the bath solution, and that, in general, the rates of potassium entry and of sodium efflux change simultaneously and in a parallel manner.

To summarize present views of the ionic maintenance mechanisms of the resting cell, one imagines that the cellular metabolism is responsible for the upkeep of ionic concentration differences between cell and surroundings, (1) by synthetizing large organic anions which cannot diffuse through the membrane, and (2) by providing the energy for an active ion-exchange mechanism, expelling sodium and accumulating potassium ions, possibly one for one. The cell membrane has a generally extremely low ionic permeability so that even in the complete absence of pumping action it will take many hours before the Na/K concentration gradients run down. The ionic permeability of the resting membrane, apart from being small, is also differentially very selective, that to potassium being much higher than to sodium, so that the PD across the resting membrane approximates to that of a potassium-concentration cell, though it does not quite reach this level (it would do so only if the sodium permeability were negligible).

The evidence for the existence of a cation pump is impressive; there is little doubt that the required Maxwell demons are being paid for; the account seems satisfactory not only in terms of available calories, but even in the form of the available currency (made up of ATP and other acceptable phosphates). But, as in most instances of specific cellular utilization of metabolic energy yields, there is no theory at present to explain how the so-called driving reaction is geared to the particular purpose of expelling sodium and piling up potassium. This is one of the familiar challenging gaps waiting to be bridged, even if only by a temporary working hypothesis.

CABLE PROPERTIES OF NERVE

It has been seen from the tracer experiments that the surface membrane presents a formidable barrier to the diffusion of ions, which agrees with the conventional picture of the membrane consisting of, or containing, a thin layer of insulating lipid material. It may correspond to the "double contours" of 50- to 100-A thickness which have been seen in electron micrographs of various cellular or intracellular surface structures.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

The conclusions which have been reached from chemical and tracer measurements are in general agreement with what is known about the electrical properties of nerve and muscle fibers. By studying alternating-current impedance, or by analyzing the attenuation and temporal distortion along the fiber surface of a small voltage-step signal applied at one point between inside and outside, it had been shown that the interior of axons and muscle fibers behaves as a conducting cylinder of slightly higher electrolytic resistivity than the outside fluid but is separated from the outside by a surface layer of very low conductance (of the order of 10^{-4} to 10^{-3} ohm^{-1} cm^{-2}) and of high capacitance. The fibers possess, therefore, electrical properties which are analogous to those of a cable, and one may ask to what extent these properties can be used for the purpose of signal transmission. Now, taken as a passive transmission line, the axon would be of little use, for its cable losses are very great: its surface leakage and the resistivity of its core are some 10^8 times greater, and its sheath capacity about 10^6 times higher than those of an ordinary commercial cable. In fact, a weak applied square-pulse signal (i.e., an electric pulse of subthreshold intensity which fails to excite the inherent relay mechanism of the axon membrane) fades out and becomes badly blunted within a few millimeters of its origin. It is clear, therefore, that the cable properties of the nerve are quite insufficient to serve the propagation of a message over the required distances, unless the deficiencies of the cable are made up by a special boosting process. This is indeed the case, and the sole purpose of the excitatory mechanism is to regenerate and to reamplify the signal at each point of the line, and so to insure its forward passage without attenuation or distortion of its brief wave form.

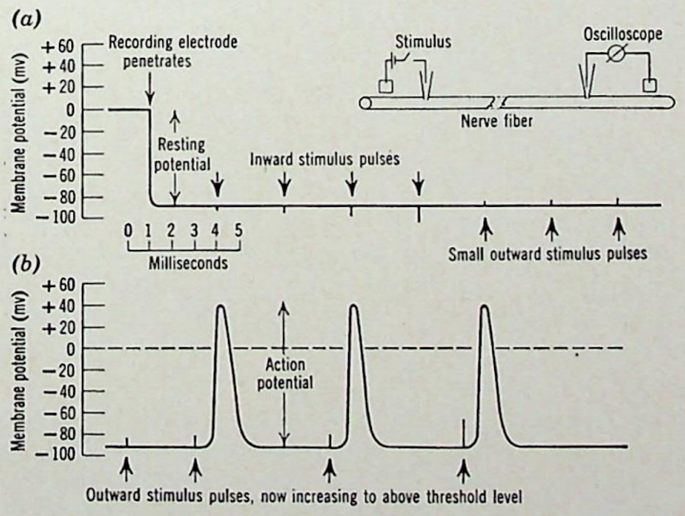


FIG. 2. Diagram illustrating the response of a nerve or muscle fiber to an electric stimulus applied at a distant point. The inset diagram shows the arrangement of stimulating and recording electrodes. In (a), a recording micropipette is inserted into the fiber interior, showing the resting potential across the cell membrane. Vertical scale indicates level of potential at the tip of the micropipette relative to the outside bath. Current pulses are applied through the distant stimulating electrodes, first inward, then outward through the membrane. No response is produced until, in (b), the outward pulses exceed a critical threshold when an all-or-none action-potential wave is recorded.

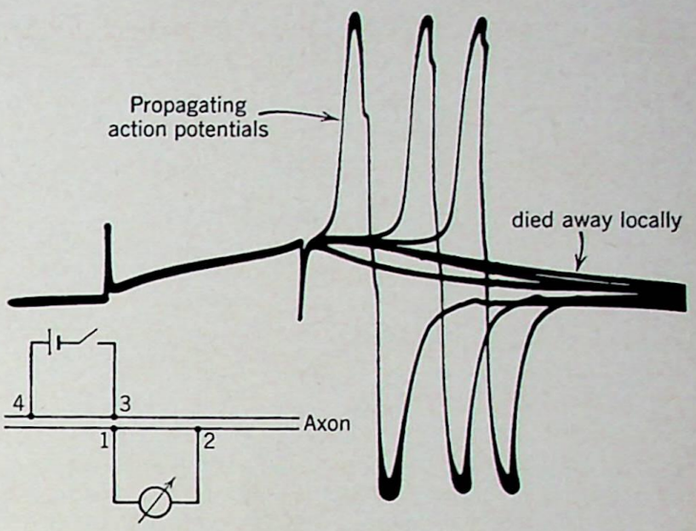


FIG. 3. Local and propagated potential changes, recorded at the site of stimulation from a single crab nerve fiber. Inset diagram shows arrangement of recording (1 and 2) and stimulating (3 and 4) electrodes, all four being placed on the axon surface, 1 and 3 close together. The records show, superimposed, the local potential changes produced by 6 successive pulses of threshold strength (duration of pulse *ca* 2 msec, beginning and end marked by brief artifacts). The potential change either died out locally or flared up after a variable delay into a large propagating action potential (about 100 mv peak-to-peak). Its diphasic nature is due to the wave of surface negativity passing first point 1, then point 2.

It is worth emphasizing that the presence of a cable-like mechanism, insufficient though it is for long-distance signaling, forms an essential link in the process of impulse propagation: it allows the action potential which has arrived at one point to impart a stimulus to the next region and to excite its latent relay mechanism. Moreover, the passive cable spread is of great importance for the integration of converging messages which takes place in the cells of the central nervous system. These cells become excited, or inhibited, by the interaction and summation of many subthreshold electrical changes which are imposed on different points of their surface membrane, all within a short range of the central cell body. The effective summation of such converging electrical influences is possible, because over short distances (a fraction of a millimeter) the cable properties of the cell are quite adequate, and the subthreshold potential changes do not suffer much attenuation. They can, therefore, sum up to the firing point at which an impulse emerges, capable of traveling the whole length of the axon process.

PROCESS OF EXCITATION

What is the nature of the excitatory relay mechanism by which the impulse is boosted to its full strength? The process can be demonstrated by recording the electrical events at the point of the fiber at which an impulse is initiated by an applied electric stimulus (Figs. 2–4). When a square pulse of current is passed inward through the axon membrane, the membrane potential is displaced

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

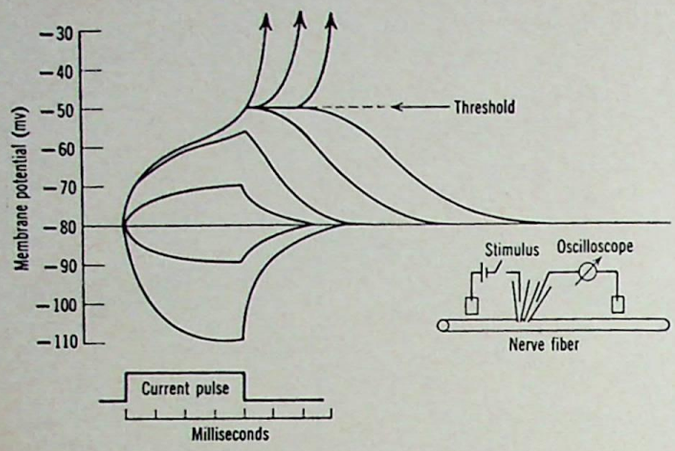


FIG. 4. Diagram showing local potential changes across axon membrane, due to rectangular current pulses (4-msec duration, varying direction and amplitude). At threshold, the membrane potential is displaced to a level at which it is in a state of unstable equilibrium, either flaring up into a propagating signal or subsiding locally.

from its resting level of, say, −60 mv inside to a higher level of internal negativity. The potential change takes a certain time to develop and to decay; this time course depends upon the resistance and capacity of the fiber, as in an ordinary leaky and capacitative cable structure. Currents of this direction, which increase the membrane PD, do not excite, but simply produce local potential changes approximately proportional to current strength, and fading-out along the fiber with a decrement of about 50% per 1 or 2 mm, characteristic of its cable coefficients.

When the polarity is reversed and weak current intensities are employed, very similar local potential changes of the opposite direction (reducing the membrane PD, or partially depolarizing) are observed. However, as the current strength is increased the picture alters. The membrane potential changes progressively, more than in proportion to the applied current, and a point is reached at which the depolarization becomes regenerative and, after withdrawal of the current, the decay of the effect greatly retarded. A slight further increase in current strength causes a flare-up of the potential change which now passes through an automatic cycle; it rises to the full height of the propagating action potential and then returns to the initial level within 1 or 2 msec. This event (known as the spike) travels at constant speed (about 25 m/sec in the squid fiber at 20°C) down the axon, without loss of amplitude anywhere; it leaves behind a short refractory period of a few msec, after which the axon is again excitable and capable of repeating the event. It is interesting that even below the firing point (or threshold), some local "smoldering" seems to go on. Indeed, there is a good analogy between the local electrical events in a nerve fiber and any other form of regenerative chain reaction, in which the initially applied stimulus is reinforced by the reaction which it produces. It would be possible, for instance, to obtain a very similar family of kinetic curves, if one used an explosive gas mixture (say hydrogen and oxygen), applying or withdrawing heat instead of current, and plotting on a suitable time scale the temperature of the gas instead of membrane potential. Well below the ignition point, the temperature changes along an exponential time course and equilibrates at a level determined by a balance between heat leakage to the surroundings and rate of heat supply. Close to the ignition point, some gas molecules combine and produce extra heat, but not at a sufficient rate to keep the reaction going when the applied heat source is switched off. At the ignition point, the rate of reactive heat production just balances the leakage to the cooler surroundings and, for a while, maintains the system in unstable balance from which either an explosive flare-up or a return to the ambient stable temperature occurs. Apart from this formal analogy, there is, of course, little resemblance between an explosive exothermic process and the much more subtle and repeatable release of electric energy by the nerve membrane.

PROPERTIES OF THE NERVE IMPULSE

Several questions arise at this point: (1) What is the nature of the regenerative factor in a nerve fiber which causes an initially applied potential change to amplify itself? (2) What brings the action potential down to the original baseline? (3) What causes the system to become temporarily inexcitable? (4) How does the axon recover from its impulse activity?

Regenerative Factor

It has been shown[16,17] that the sodium conductance of the membrane depends upon the electric field across it; the sodium conductance is very low at the normal level of the resting potential, but increases, temporarily, when this potential difference is reduced. *How* this comes about is not known and remains one of the outstanding problems, but the statement that sodium permeability is raised by a depolarization of the axon membrane is based on very strong evidence. The consequences of this effect are far-reaching: suppose a partial depolarization is produced by passing current outward through the membrane. The consequent rise in sodium permeability causes an increased leakage of these ions down their concentration gradient into the fiber. As a result, positive change is transferred to the interior and the membrane potential is diminished further. This is the regenerative step which causes the permeability to sodium to rise again and more sodium ions to enter. At a certain level of the membrane potential, the threshold, this process becomes progressive and flares up into the action-potential wave. Like the ignition point of the explosive-gas analogy, threshold is a point of unstable equilibrium at which restoring and regenerative tendencies just balance each other. The regenerative factor responsible for the automatic ascent of the action potential is the progressive relation between sodium-conduct-

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

ance change and fall of membrane potential; the restoring factor is to be found in the potassium (and chloride) conductances which tend to reset the membrane potential to the initial stable baseline.

Once the threshold level is exceeded, the sodium conductance becomes rapidly the dominant factor and the regenerative entry of sodium would continue until the membrane potential swings well to the other side, and the interior attains sufficient positive charge to balance the inward concentration gradient and prevent any further net influx of sodium ions. In other words, if the initial excitatory process continued, a new equilibrium level would be reached corresponding to that of a sodium-concentration cell. This is, indeed, the accepted explanation for the reversal of the membrane potential during the peak of the impulse. The level which is attained falls somewhat short of the sodium potential, because the initial permeability change is only transient: within about a millisecond, the high, specific sodium conductance is switched off and apparently is converted into a state of high potassium conductance.

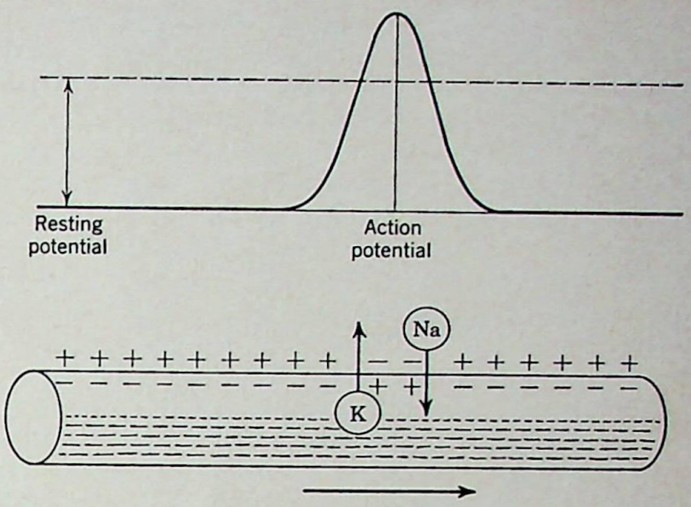


FIG. 5. Diagram of action potential traveling (in direction of arrow) along the axon. During the rise of the wave, sodium ions enter the fiber and charge the interior positively. During the decline, potassium ions leave the fiber and restore the initial membrane potential.

Restoring Factor

This secondary changeover, from sodium to potassium permeability, ensures a rapid return of the potential to its original baseline, for potassium ions now will leave the fiber rapidly in the direction of their electrochemical gradient. Such electric restoration would be effected even without a secondary increase of potassium conductance, provided the initial sodium-permeability rise is cut off (a process known as inactivation). In this event, however, the decline of the action potential would take a relatively long time and proceed with the time constant of the resting-fiber membrane. In the squid axon, it would take several milliseconds at the ordinary leakage rates of potassium (and chloride) to restore the membrane potential to the resting level. By opening up the potassium channels, this process becomes greatly accelerated, and a few milliseconds later the axon is ready to fire again.

The work of Hodgkin and Huxley has shown that a depolarization imposed on the axon membrane (by the so-called voltage-clamp method) produces these two successive permeability changes (Na, then K), the intensity of both changes increasing similarly with the amplitude of the enforced voltage step. During the normal action potential, the sodium-conductance rise is regenerative and self-reinforcing, while the potassium-permeability change is restorative and gradually cuts itself down as the membrane potential is caused to return to the original stable level.

It should be noted that the electrical return of the system during the impulse is *not* brought about by a reversal of the ionic movement (which would mean a forced expulsion of the quantity of sodium which had entered during the ascending phase of the action potential), but by the leakage of an equivalent quantity of potassium ions (see diagram in Fig. 5). Both sodium and potassium are moving downhill during the impulse; the quantities, however, are so small that no detectable change of the internal concentration would occur during one impulse, and there is plenty of time for the pumping process to restore the chemical situation in resting periods later on.

Nature of the Refractory Period

The secondary permeability effects which bring about the decline of the action potential are also responsible for the short period of inexcitability (refractoriness) which follows the propagation of each impulse. Clearly, during the state of inactivation—i.e., while the regenerative dependance of sodium permeability on membrane potential is in abeyance—an applied electric current fails to excite; secondly, a state of very high potassium permeability implies that the restoring factor is greatly strengthened and tends to oppose any displacement of the membrane potential from the stable potassium-equilibrium level. Hodgkin and Huxley have determined the rate constants of these changes and their time course of decline following a restoration of the normal membrane potential. It was found that the period of recovery of excitability after the impulse is simply a measure of the time by which the return of the normal permeability relations lags behind the return of the membrane potential.

Chemical Recovery

There are several phases of restoration: the recovery of the membrane potential is followed by a return of excitability, but the system has not yet reverted completely to its initial state; in fact, a minute fraction of the intracellular potassium store has been lost and replaced by sodium. The downhill movement of sodium and potassium ions across the membrane has provided

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

the energy for the nerve signal. The amount of ions which exchange during the impulse is extremely small; it has been determined for a series of impulses in tracer experiments, and by chemical analysis of single fibers, and was found to exceed the minimum coulomb requirement by a factor of 3. The transfer of charge needed to alter the membrane potential by 100 mv, with a capacity of 1 μf cm^{-2} is 10^{-7} coul which requires a net transfer of about 10^{-12} M of univalent ions: the net exchange of sodium and potassium per impulse is about 3 to 4 times as much.[16] These quantities produce only a minute internal concentration change per single impulse (about 0.0001% in the large squid axon, but considerably more in small fibers whose volume/surface ratio is correspondingly less). Even if the ion pump were not active, a squid axon could conduct some hundred thousands of signals before its accumulated ion store would be exhausted. As in many other forms of brief, high-intensity biological action, a period of impulsive activity is completed at the expense of small but cumulative chemical losses; these are made good during periods of rest by continuous chemical recovery mechanisms which Connelly discusses (p. 475).

A rough qualitative picture has been given here of the conclusions derived from the extremely accurate quantitative work of Hodgkin and Huxley and others, and reference is made to their papers for a description of the experimental evidence on which these conclusions are based.[17–20] Summarizing briefly, it has been found that the resting potential is related to the external potassium but independent of sodium concentration, while the peak of the action potential depends on the sodium-concentration ratio in the way predicted by the Nernst theory. In experiments in which the axon membrane was subjected to a voltage clamp, it was found that a sudden stepwise depolarization (e.g., to zero membrane potential) was followed by a surge of positive inward current followed by a maintained outward current. The first, transient, component (in which current flows "in opposition to Ohm's law" indicating the presence of the required negative resistance, or regenerative factor) could be identified as owing to the inrush of sodium, for it could be made to reverse in polarity by reversing the concentration gradient of sodium ions. With a variety of sodium concentrations, the reversal was shown to occur invariably at the electrochemical equilibrium potential for sodium ions. The second, maintained, component was identified by tracer-efflux measurements as a maintained high-intensity outward current of potassium ions. The rate constants of the several successive conductance changes were determined, and from this information the properties of the impulse, its excitatory threshold, its amplitude, duration, and conduction velocity, could be synthetized with a high degree of precision.

While this series of investigations has been brought to a decisive conclusion, this should, of course, not conceal the fact that the nature of the permeability relations and the reason why membrane potential and ionic conductances are linked in such a specific manner remains an altogether unsolved problem at the present time.

SOME SPECIAL CASES

Medullated Nerve Fibers

The picture presented so far describes the situation in the giant axon of the squid, on which most of the crucial experiments have been carried out. But substantially the same conclusions apply to all other types of nerve fibers and to most types of skeletal-muscle fibers which have been studied. There is, however, an important structural difference between the so-called nonmedullated nerves, and the medullated (myelinated) axons of vertebrates which possess relatively thick, segmented sleeves of myelin broken at regularly spaced intervals (the nodes of Ranvier about one or a few millimeters apart) at which a small area of axon membrane is exposed to the outside fluid. In the ordinary nonmedullated fibers, the propagation of the signal depends upon the existence of a continuous relay mechanism which is built in along the whole line at many scattered random points of the axon membrane. This makes up for the deficiencies of the poor cable transmission: without it, the action potential would be attenuated rapidly in the region ahead. Experimentally, this can be shown to be the case by poisoning or anaesthetizing a narrow region of the axon cylinder and so abolishing locally the regenerative mechanism. Provided, however, the attenuated signal which spreads to the other side of the blocked region exceeds the threshold depolarization (which amounts to only $\frac{1}{10}$ to $\frac{1}{5}$ of the full amplitude of the spike), the relay process will start again and the impulse will jump the block.

In medullated nerves, transmission of signals normally occurs in jumps.[21,22] The segmented myelin sheath provides a relatively good, low-capacity insulation around the axon, and the relay mechanism is concentrated and restricted to the discrete nodes of Ranvier where the axon membrane is exposed. There is, in other words, a rather close miniature analogy to a submarine cable with relay stations interposed at the nodes. In the internodal region, the fiber behaves as a relatively good cable with only slight capacitative losses which are made good at the next node. The final result is a great improvement in speed and economy of signaling: the ionic losses are restricted to relatively few microscopic areas of the fiber surface, and the propagation velocity of the impulse is some 10 times greater than for a nonmedullated axon of the same size. To achieve the speed which medullated axons attain *without* the provision of a myelin sheath, the only alternative way for nature would have been to reduce the internal core resistance; that is, to increase the fiber size. Considering the need for a large number of separate signaling channels within a confined space (there are one million medullated axons

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

in the human optic nerve alone), the great advantages of myelination are obvious.

Other Tissues

Most of the tissues, nerve or skeletal muscle, whose function depends upon the rapid propagation of an all-or-none impulse, appear to make use of a regenerative entry of sodium ions in the course of excitation; there is, however, some doubt as to whether the secondary increase of potassium permeability is a similarly wide-spread phenomenon, or whether it may be a more or less dispensable effect. In some tissues, e.g., certain medullated axons, the potassium resistance and the electric time constant of the resting membrane appear to be sufficiently low to insure a rapid decline of the action potential once the sodium conductance has dropped back towards the initial low level.

In heart muscle, the action potential has a different shape, of "flat top" and long duration. The initial fast rise is probably governed by a sodium mechanism just as in nerve, but the subsequent changes differ at least in quantity. It may be that the maintained plateau of depolarization arises from the increased sodium conductance not being switched off completely, and from a failure of the potassium efflux to rise above its steady rate. It is interesting to consider the usefulness of this long flat-topped heart potential: its duration apparently governs the systolic contraction of the heart muscle, and because of the relatively low frequency of its regular beat, there is no need for rapid electric recovery of the membrane potential.

In skeletal muscle, a quantitative difference from nerve was found in the large capacity of the surface membrane (5 to 10 $\mu f/cm^2$ as against 1 $\mu f/cm^2$ in nerve). There is a correspondingly larger transfer of charge during the action potential, and recent tracer measurements by Hodgkin and Horowicz[23] have shown that the amounts, per unit surface area, of sodium and potassium exchange are about 3 times greater than in the nerve fibers. There appears to be less temporal overlap of the two ionic permeability changes; indeed, resistance measurements indicate that there are two discrete phases of high membrane conductance during the action potential. This would enable the sodium ions which enter to produce a relatively larger net effect. It also seems that the delayed rise of potassium conductance is shortlived and stops before the action potential has returned to the baseline, giving rise to a characteristic long tail (the so-called negative afterpotential) of the muscle spike.

There are finally some tissues (e.g., crustacean muscle and certain conducting plant cells) in which action potentials of the usual propagating all-or-none type can be elicited by electric stimulation, but sodium entry does not seem to be responsible for their regenerative ascent. In many of their features, these action potentials resemble those of the nerve axon very closely. It appears that the underlying events are similar and involve ionic permeability changes, but that the specific channels which are being opened during the depolarization of the membrane are different. For example, in certain algae, a regenerative exit of chloride from the cytoplasm takes the place of sodium entry,[24] while in crustacean muscle, entry of divalent cations possibly may be responsible for the rise of the action potential.[25]

Structure and Function

To return to some of the problems raised by F. O. Schmitt (p. 455): Can one try now to identify our physiological mechanisms with any one of the structures revealed by the electron microscope? The question has been raised, for instance, whether ion permeabilities are entirely to be attributed to the axon membrane, or could they be properties of Schwann cells which envelope the axon? Insofar as the observations have been repeated on skeletal-muscle fibers, which are not invested with satellite cells, and insofar as they corroborate the main conclusions derived from the nerve experiments, one may feel that the presence of a Schwann-cell envelope does little to mask or mimic the electrical properties of the axon membrane. It is important also to recall that the cellular investment of the axon is not complete, but there are gaps through which the axon surface communicates with the extracellular fluid. Although these channels may be narrow (only a few hundred Ångström units wide), they are also short (a few microns long) and probably add only a small resistance in series with the axon membrane. There is evidence for such an external nonreactive resistance of about 5 ohms in series with a resting membrane of approximately 1000 ohms $\times cm^2$. Recent work of Frankenhaeuser and Hodgkin[26] suggests that the presence of this external bottleneck gives rise to a slow accumulation of potassium ions on the surface of the squid axon during a period of prolonged impulse activity. This is associated with cumulative changes in the membrane potential, but these are rather minute by comparison with the main features of the spike. Nevertheless, one must bear in mind the possibility that the distribution of current flowing through the axon membrane may be modified locally by Schwann cells which cling to its surface and by the presence of discrete channels between them.

It may seem a little disappointing that, with all of the high-resolving power which has been attained in recent years, both in the study of fine structure and of the functional mechanism of nerve, there is still such a conspicuous lack of correlation between these two fields of work. One of the reasons is undoubtedly that the details which are revealed in their most striking form by such different techniques are not necessarily related to the same features. In the electron micrograph, the almost complete investment of an axon by a Schwann cell, or the almost intimate contact between the two surfaces, may be the most impressive phenomenon,

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

while from the point of view of electric current distribution the appearance of a fine gap or crack in this investing layer (which is difficult to find and might easily be discarded as an artifact) may be a much more important feature.

A second reason for the apparent discrepancy is that there must be structural counterparts of functional aspects other than those with which the student of the nerve impulse is directly concerned. The purpose of this paper is to describe some efforts that have been made to elucidate the nature of the nerve signal. But, in addition, there is built into the nerve a machinery for chemical recovery processes, and for the maintenance of the axon and its enzymic apparatus, which one associates with the so-called trophic influence of the nerve-cell body and its nucleus. Little is known about this process, but there are many indications that satellite cells are actively involved, both in trophic maintenance and in the disintegrative processes which follow the severance of the axon from the cell body.

The elementary membrane structure which has been seen in electron micrographs, e.g., of permanganate-fixed axons, is a double contoured-layer 50 to 100 A thick. It is not known what fraction of this layer is composed of the insulating lipid material which we hold responsible for the ion-impermeable and capacitive properties of the axon membrane. Nor can one be certain that improvements in microscopic resolution will not show further subdivisions and even thinner membrane elements. And ultimately, the physiologist is interested not so much in the lipid bulk phase of the membrane which occupies most of it, and which seems never to change nor to take part in the transport of ions, but in those very sparse molecular patches—possibly at continuously varying sites, at which ionic gaps or channels are thought to be formed, and where "carrier molecules" are believed to be active in transferring sodium and potassium ions from one side to the other.

BIBLIOGRAPHY

[1] E. Du Bois-Reymond, *Untersuchungen über thierische Elektricität* (G. Reimer, Berlin, 1848–1860), Vols. I, II.

[2] L. Hermann, *Untersuchungen zur Physiologie der Muskeln und Nerven* (Hirschwald, Berlin, 1868).

[3] J. Bernstein, Pflüger's Arch. ges. Physiol. **92**, 521 (1902).

[4] A. L. Hodgkin, Proc. Roy. Soc. (London) **B148**, 1 (1958).

[5] A. L. Hodgkin, Biol. Revs. Cambridge Phil. Soc. **26**, 339 (1951).

[6] A. F. Huxley in *Ion Transport across Membranes*, H. T. Clarke, editor (Academic Press, Inc., New York, 1954), p. 23.

[7] B. Katz, Symposia Soc. Exptl. Biol. **6**, 16 (1952).

[8] A. L. Hodgkin and R. D. Keynes, J. Physiol. **119**, 513 (1953).

[9] K. S. Cole and H. J. Curtis, J. Gen. Physiol. **22**, 649 (1939).

[10] A. L. Hodgkin and A. F. Huxley, Nature **144**, 710 (1939).

[11] H. J. Curtis and K. S. Cole, J. Cellular Comp. Physiol. **19**, 135 (1942).

[12] P. J. Boyle and E. J. Conway, J. Physiol. **100**, 1 (1941).

[13] R. D. Keynes, J. Physiol. **114**, 119 (1951).

[14] A. L. Hodgkin and R. D. Keynes, J. Physiol. **128**, 28 (1955).

[15] P. C. Caldwell and R. D. Keynes, J. Physiol. **137**, 12P (1957).

[16] R. D. Keynes and P. R. Lewis, J. Physiol. **114**, 151 (1951).

[17] A. L. Hodgkin and A. F. Huxley, J. Physiol. **116**, 449, 473, 497 (1952).

[18] A. L. Hodgkin and A. F. Huxley, J. Physiol. **117**, 500 (1952).

[19] A. L. Hodgkin and A. F. Huxley, J. Physiol. **121**, 403 (1953).

[20] A. L. Hodgkin, A. F. Huxley, and B. Katz, J. Physiol. **116**, 424 (1952).

[21] A. F. Huxley and R. Stämpfli, J. Physiol. **108**, 315 (1949).

[22] I. Tasaki, *Nervous Transmission* (Charles C. Thomas, Springfield, 1953).

[23] A. L. Hodgkin and P. Horowicz, "Movements of Na and K in single muscle fibres," J. Physiol. (to be published).

[24] C. T. Gaffey and L. J. Mullins, "Ion fluxes during the action potential in *Chara*," J. Physiol. (to be published).

[25] P. Fatt and B. L. Ginsborg, J. Physiol. **142**, 516 (1958).

[26] B. Frankenhaeuser and A. L. Hodgkin, J. Physiol. **131**, 341 (1956).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

51
Recovery Processes and Metabolism of Nerve

C. M. CONNELLY

The Rockefeller Institute, New York 21, New York

INTRODUCTION

THE prime function of peripheral nerve is the conduction of impulses. In biological terms, the nerve impulse is the means of transmission of an item of information from one point in an organism to another. In terms of physics and chemistry, the impulse is a spatially propagating, transient disturbance in the state of a complex reaction system; it might be detected by, or it might be described in terms of, any of those conceivably measurable changes which are a part of it or to which it gives rise. In thermodynamic terms, the impulse is a dissipative process in which the free energy of the nerve and its environment is decreased, principally by a flow of ions down gradients of their electrochemical potential. In this discussion, attention is focused upon processes by which resting nerve maintains itself in a steady state, ready to function, and upon processes which serve to restore nerve to its resting state after it has conducted impulses. A central question is how energy-yielding processes may be coupled to those requiring energy. In an approach to a more specific formulation of this question, the discussion emphasizes correlations between changes in rate of oxidative metabolism and electrochemical manifestations of ionic movement.

In resting nerve there are at least two processes which may be presumed to require energy. These are the maintenance of structural integrity and the maintenance of the ionic-distribution characteristic of the resting state. About the former very little can be said. Perhaps the cell bodies from which the excised nerve fibers have been severed are the primary site of synthetic reactions underlying such maintenance. During activity and recovery therefrom energy demand may be increased in two ways: there must be an acceleration of ion transport processes in order to reverse the exchange of ions which occurs during the passage of impulses and, secondly, there may be a dissipation of chemical energy associated with the permeability cycle that allows the ionic exchange to take place.

Insofar as nerve at rest is in a steady state and following activity returns to the same steady state, the over-all process of impulse conduction involves no net external work and results only in the conversion of chemical energy to heat.

The following discussion indicates that oxidative metabolism in nerve can serve as an adequate energy source. Brink[1] has recently reviewed the evidence that the biochemistry of peripheral nerve of the frog resembles that of other animal cells; the nerve appears to contain and utilize the Meyerhof-Embden pathway of carbohydrate breakdown, the Krebs tricarboxylic acid cycle, and the cytochrome chain of enzymes. The energy turnover appears to be mediated by the usual system of phosphate compounds (adenosine phosphates and creatine phosphate) which are replenished principally by oxidative phosphorylation carried out by mitochondria. In what follows, it will be assumed (1) that any processes which rely upon metabolic energy do so by chemical reactions coupled to the breakdown of adenosine triphosphate, or its equivalent, and thereby increase the intracellular level of phosphate acceptor; (2) that the kinetics of oxygen uptake by nerve are some reflection or measure of the changes in concentration of phosphate acceptors at the mitochondria. A particularly dramatic experiment directly supporting the first of these assumptions is that in which Caldwell and Keynes[2] injected ATP into a metabolically poisoned squid axon and observed a partial restoration of the rate of sodium extrusion.

Firstly, the kinetics of the increases in oxygen utilization associated with the conduction of impulses is described and compared with the heat measurements of Hill, and then it is shown how the kinetics are modified under circumstances in which rates of ion transport have presumably been affected. Secondly, data on the ionic fluxes across the nerve surface are examined and the estimated energy requirements of transport processes are compared with energy available from oxidative metabolism. Thirdly, some observations are described of prolonged positive afterpotential (post-tetanic hyperpolarization) which appears to be closely related to ionic transport processes.

Most of the observations which follow come from experiments on the excised sciatic nerve of the frog. This preparation has some advantages over the giant axon of the squid in that its metabolic and electrical properties remain more nearly constant over an experimental period of ten to fifteen hours. It has, however, the disadvantage of being a bundle of fibers of several different types and there are complications arising from the existence of an appreciable extracellular space. The minimum environment required by frog nerve to maintain function and a satisfactory steady state at 20°C is a balanced bathing solution containing sodium, potassium and calcium salts, and dissolved oxygen. It should be emphasized that many or most of the electrical and metabolic properties of nerves appear to be basically similar from one species of animal to another, and similar also, in fact, to the properties of other excitable tissues such as skeletal and cardiac muscle.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

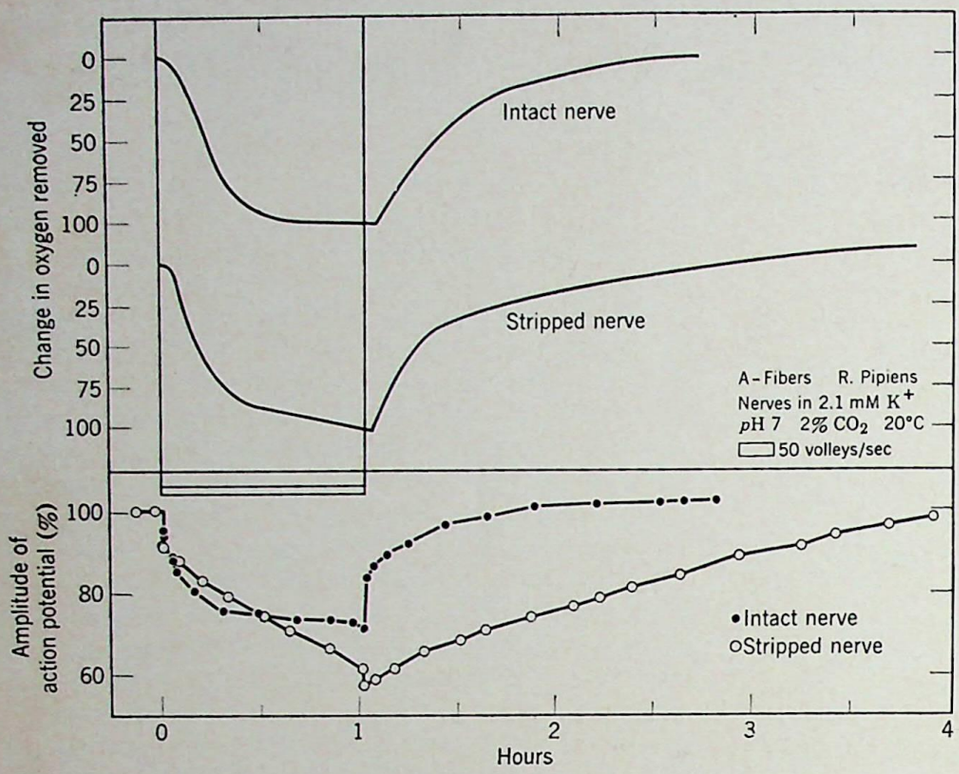


FIG. 1. (Upper) Relative time courses of the increases in oxygen uptake resulting from activity in intact frog sciatic nerve (perineurium not removed) and in stripped nerve (perineurium removed). Measurements made in a flow respirometer[3] with a polarized oxygen cathode. Polarographic current is a linear measure of dissolved oxygen remaining in solution after it has flowed past the nerve, a downward deflection in the trace, as shown, indicating an increase in oxygen uptake. (In all tetani in this and the following figures, only type A fibers stimulated.) (Lower) Time courses of changes in amplitudes of compound action potentials of the two nerves, measured during tetanus and occasionally by test shocks during recovery.

OXIDATIVE METABOLISM AND HEAT PRODUCTION

The oxygen uptake by resting frog nerve is usually 30 to 40 $mm^3/g(wet)/hr$ or about 1.5 μmoles/g(wet)/hr. When a nerve is tetanized, its rate of oxygen uptake increases and approaches a new steady level in about 30 min, closely following an exponential time course with a time constant of 5 to 8 min, Fig. 1, upper curve.[3] The amplitude of the increase is greater the higher the frequency of tetanus, at low frequencies, and approaches a maximum limiting value (of about 1 μmole/g (wet)/hr) at about 100 volleys/sec,[4] as shown in Fig. 2. At the end of tetanus, the rate of uptake declines slowly toward the resting value along an exponential whose time constant is usually 15 to 25 min. [After tetani at frequencies lower than about 10/sec, recoveries may be more rapid than this (see Fig. 11 in an article by Connelly *et al.*[5]).]

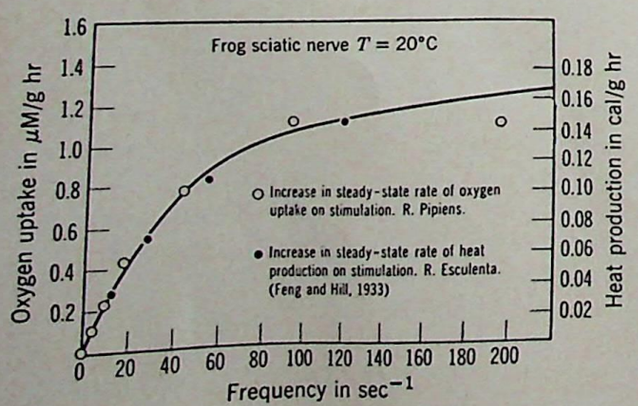


FIG. 2. Oxygen uptake and heat output for active nerve in relation to frequency of conducted impulses. Perineuria not removed. Ordinate scale symbol μM means micromole [from F. Brink, D. W. Bronk, F. D. Carlson, and C. M. Connelly, Cold Spring Harbor Symposia 17, 53 (1952)].

The over-all increases in heat production by tetanized frog nerve show kinetics similar to those of the increases in oxygen consumption.[6] The quantitative comparison of steady-state heat rate and oxygen consumption as a function of frequency is shown in Fig. 2. The ratio of the right-hand to the left-hand scale is within 15% of the accepted value of the calorific equivalent of oxygen, 5 cal/cc. The agreement is quite satisfactory, in view of the difference in species.

A detailed analysis of the heat production of nerve during short tetani reveals the presence of a component whose onset and termination are abrupt and correspond to the beginning and end of the tetanus.[7] This "initial heat," only a few percent of the total heat associated with the tetanus, appears to have no counterpart in the time course of the increase in oxygen uptake, the curve of which rises linearly from the beginning of the tetanus.[4] In recent elegant work, Hill and his coworkers[8] have resolved the initial heat associated with a single volley in crab nerve into a positive phase (heat production) followed by a smaller negative phase (heat absorption). These phases may result from the heats of dilution or mixing of the sodium and potassium ions exchanged during the impulse, but other events of the permeability cycle have not been excluded as contributing causes.

The second curve of Fig. 1 illustrates the time course of increase in rate of oxygen uptake by a nerve trunk from which the perineurium, or connective-tissue sheath, has been removed. The striking difference between this and the upper curve is that the recovery is characterized by two components, the first lasting no more than about 30 min and the second extending over several hours.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

The lower part of Fig. 1 shows the accompanying changes in the amplitudes of the compound action potentials of the two nerves, the stripped nerve showing much less rapid recovery than the unstripped nerve. The possibility that these differences in the metabolic and electrical behavior of stripped and intact nerve might result principally from differences in the extracellular level of potassium ions was tested by the series of experiments illustrated in Figs. 3 and 4. These compare the kinetics of the changes in oxygen uptake and action potentials, respectively, of stripped nerves bathed in solutions containing different concentrations of potassium ion. The three lower curves of Fig. 3 possess fast components of about the same relative magnitude and time course whereas the slow components of these curves show gradation from effectively no recovery in K-free solution, to a slow almost linear recovery in 2.1 mM K^+, to a somewhat more rapid curvature characteristic of an exponential of about a 40-min time constant, in 5 mM K^+. In the top curve (8.5 mM K^+), the over-all recovery is even more rapid than any observed in intact nerve (in 2.1 mM K^+), with an exponential time constant of only 11 min instead of 15 to 25. It is not possible to distinguish two phases of recovery in this case and it is not clear whether the slower phase observed in the lower curves has been speeded up to merge with the rapid phase or whether it has decreased in magnitude to zero. In Fig. 4, one sees that the lower the concentration of potassium in the bathing solution, the more rapid is the decline in height of the action potential during tetanus and the less rapid is its recovery. Recovery of action potential is incomplete in K-free solution, as the recovery of oxygen uptake was seen to be. In 8.5 mM K^+, however, not only does the action potential not decline in amplitude during tetanus but it shows about 13% increase in amplitude during the post-tetanic period.

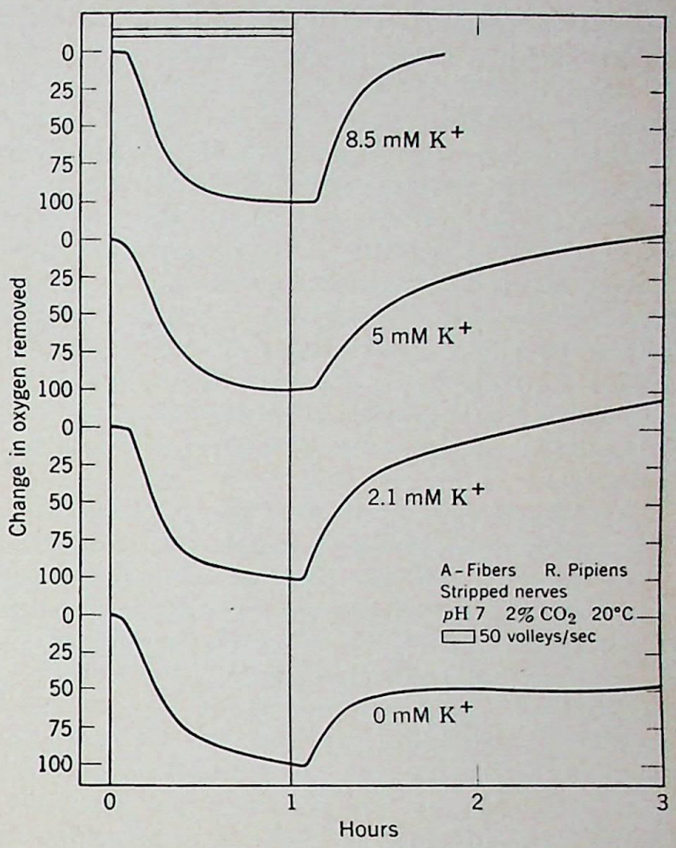


FIG. 3. Relative time courses of the increase in oxygen uptake by stripped nerve bathed in Ringer's solutions containing different concentrations of K^+. Ordinate as described for Fig. 1.

A possible key to the interpretation of these observations is the finding of Hodgkin and Keynes[9] (see also Hodgkin's Croonian Lecture[10] for an extended discussion of the movements of ions in giant nerve fibers) that the efflux of sodium from *Sepia* fibers is increased in a solution containing more potassium than does sea water and is depressed by K-free solution to one-third or one-

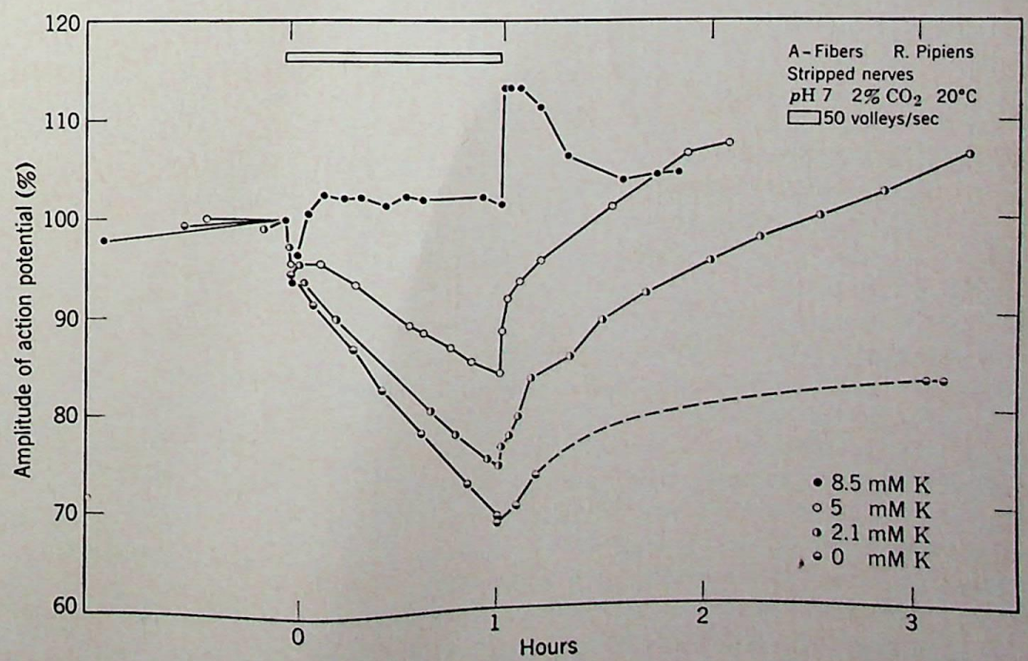


FIG. 4. Time courses of changes in amplitudes of compound action potentials of the four nerves of Fig. 3.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

quarter of the efflux into sea water. If, in frog nerve, the rate of sodium extrusion is similarly dependent upon the level of potassium in the external solution, the slow component of recovery respiration may be interpreted as reflecting the energy demand of the transport machinery. The fast component, on the other hand, might result either from cessation at the end of tetanus of energy demand by the excitability cycle, or from a rapid reduction in rate of ion transport, at the end of tetanus, because of relaxation by diffusion of the internal ionic gradients maintained near the nerve surface by the entrance of sodium and by the exit of potassium during each impulse. The following considerations suggest that additional factors may affect the kinetics of the fast component.

The principal locus of ionic flow during an impulse in a myelinated fiber appears to be across the membrane at the node.[11,12] If ionic recovery takes place principally across the nodal membrane, the question arises as to whether or not phosphate compounds participating in this recovery are rephosphorylated by mitochondria close to the node or by those farther down the internodal space. Thus, longitudinal diffusion of phosphate acceptors in the internode may determine in part the kinetics of changes in rate of oxygen uptake. These considerations apply, as well, to the kinetics of the rising phase of uptake that occurs during the tetanus. The fact that the time constant of this rising phase in intact nerve is as long as it is (5 to 8 min) and does not vary markedly with tetanus frequency[4] is consistent with the idea that the kinetics of this phase are restricted in part by configurational parameters.

The differences in the kinetics of the changes in oxygen uptake and in the action potentials of intact and stripped nerves in 2.1 mM K^+ Ringer's appear, then, to be due to restrictions imposed by the perineurium on the diffusion of ions between the extracellular space and the bathing solution. The differences arise from the secondary effects of a changing ionic composition around the fibers of the intact nerve, especially as regards the concentration of potassium.

A most important and revealing experiment is that of measuring the respiration of nerve bathed by a solution in which lithium has replaced sodium ion. It is known that lithium can substitute for sodium in the conduction process. As soon as lithium Ringer's is substituted for sodium Ringer's, the resting respiration begins a slow decline. Tetanization produces almost no increase in rate; there may result a slowing or interruption in the decline of the resting respiration, but over all the effect of activity during exposure to lithium Ringer's is certainly not more than 5% of the effect measured in sodium Ringer's. The observation suggests strongly that lithium ions are extruded from frog nerve very slowly or not at all. Swan and Keynes[13] have reported that lithium ions are extruded from frog muscle much less rapidly than are sodium ions. Another conclusion from this experiment is that energy demand originating in the excitability cycle is less than 5% of the total. Therefore, it appears unlikely that the fast recovery components of the three lower curves of Fig. 3 result from the cessation of such a demand.

ION TRANSPORT AND ENERGY REQUIREMENTS

In Table I are shown data from the work of Hurlbut and Asano on water distribution, ion distribution and fluxes, and the net ion changes which result from activity or exposure to an altered environment. The water distribution refers to intact nerve. The assignment of water and ions is based upon the assumption that the nerve consists of two compartments, intracellular and extracellular, the latter occupied by bathing solution. The internal concentrations of ions are based upon measurements in which the extracellular space had been washed free of sodium and potassium ions and include a small correction for sodium lost during this washing.[14] The fluxes of sodium and of potassium were measured as outfluxes from resting nerve which had previously been equilibrated with radioactive ions.

At the bottom of Table I the sodium gained and the potassium lost during activity and during exposure to oxygen-free and K-free environments are compared. It is apparent that in each case the sodium gain is nearly equal to the potassium loss. Asano and Hurlbut[15] have shown that ionic recovery does take place in the three or four hours following activity (50 volleys/sec, 1 hr, in 2 mM K^+ Ringer's). Figure 5 indicates that recovery from rather large ionic shifts takes place in such a way that the potassium movement in one direction balances

TABLE I. Frog nerve.

Extracellular water content	45% wet weight[a]
Intracellular water content	29% wet weight[a]
Dry weight	26% wet weight[a]

	Concentration: Intracellular $\left(\frac{\text{mmole}}{\text{kg intracell water}}\right)$	Concentration: Extracellular (mM)	Flux $\left(\frac{\text{mmole}}{\text{kg intracell water, hr}}\right)$
Sodium	47[a]	116	23[b]
Potassium	159[a]	2	10–28[b]

	Net changes resulting from: Activity 50 volleys/sec 1 hr	Asphyxia 5 hr	K-free Ringer's 5 hr
Sodium gain $\left(\frac{\text{mmole}}{\text{kg intracell water}}\right)$	18[c]	35[a]	31[b]
Potassium loss $\left(\frac{\text{mmole}}{\text{kg intracell water}}\right)$	21[c]	41[a]	34[b]

[a] Reference 14.
[b] W. P. Hurlbut and T. Asano (unpublished observations).
[c] Reference 15.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

the sodium movement in the other. The rates of ionic recovery during the first hour or two (Fig. 5) produce concentration changes of about 12 mM (intracellular) per hour. This net ionic exchange that is measured during recovery is presumably superimposed upon the normal fluxes, measured in resting nerve, of about 23 mM (intracellular) per hour for sodium and probably about the same for potassium.

For a two-compartment system, the thermodynamic expression for the energy required to move one mole of sodium from inside (i) to outside (o) and one mole of potassium in the opposite direction is

$$RT\left(\ln\frac{Na_o}{Na_i}+\ln\frac{K_i}{K_o}\right),$$

where Na_o, . . . etc., are ion concentrations (strictly activities). This expression has a value of about 3000 cal/mole for the concentrations of sodium and potassium given in Table I. If the energy available from the hydrolysis of one mole of ATP to ADP, under intracellular conditions, is 7000 to 12 000 cal, then energetically it should be possible for some 2 to 4 sodium ions to be extruded for each molecule of ADP produced. Assuming, conservatively, that one ADP is produced for each sodium ion transported, it may then be asked whether the observed oxygen consumption appears to be capable of providing sufficient energy, via phosphorylated intermediates, to transport ions at the rates observed. For frog nerve at rest, the oxygen uptake is about 1.5 μmole/g (wet)/hr and the sodium flux about 6.6 μmole/g (wet)/hr. Thus the ratio Na:O is $6.6/(2\times1.5)=2.2$. If, in the functioning cell, mitochondria maintain an average P:O ratio (phosphate acceptors phosphorylated to oxygen atoms reduced) of about 3, as isolated mitochondria do,[16] then it would seem that oxidation provides adequate energy for transporting one sodium ion per ATP broken down. If the Na:P ratio is actually unity, the comparison implies that only about 30% of the resting respiration is available for processes other than ion transport.

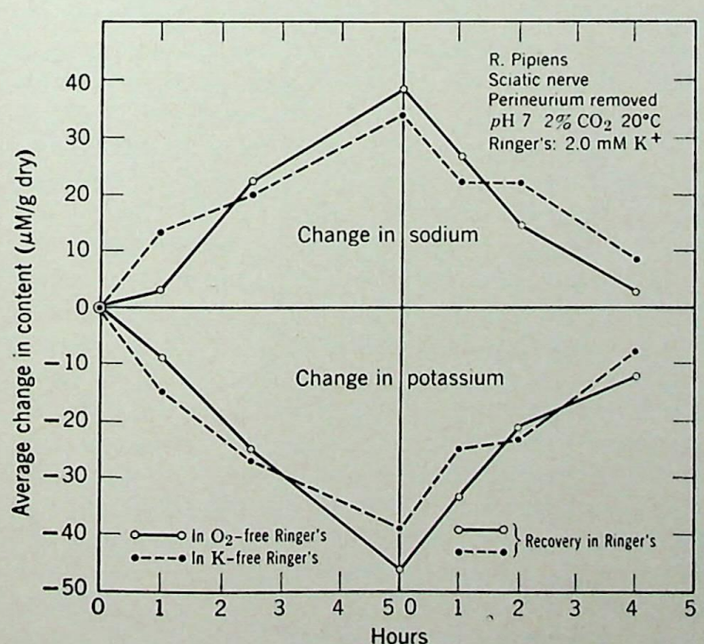


FIG. 5. Time courses of the changes in ionic contents of stripped nerves during and after exposure to oxygen free or potassium-free Ringer's solution. (Partially unpublished observations of W. P. Hurlbut and T. Asano.) Ordinate scale symbol μM means micromole.

An estimate of the increase in rate of sodium extrusion during a period of activity can be made from other observations of Asano and Hurlbut.[15] They found that after one hour of activity at 50 volleys/sec the net gain of sodium averaged 32 μmole/g (dry) when the nerve was bathed with K-free solution and only 16 when the Ringer's contained 5 mM K^+. This is an equivalent difference in rate of 4.2 μmole/g (wet)/hr. If, in the two solutions, about the same amount of sodium entered the nerve during each impulse, the observed difference may be attributed to effective, restorative transport taking place during activity. The ratio of this difference in rate to the increase in respiration during activity in 5 mM K^+ is approximately $Na:O=4.2/(2\times.8)=2.6$. This figure suggests that, in the case of active nerve also, there may be little energy utilized for purposes other than ion transport.

It should be emphasized that only one plausible line of thought has been followed in making these estimates and comparisons and that assumptions without direct experimental support have been invoked. Alternative lines of reasoning may ultimately prove to be more acceptable than the one outlined here.

POST-TETANIC HYPERPOLARIZATION

Experiments which indicate that there are measurable electrical changes associated with the ionic transport events of recovery[17] are now described.

A stripped frog nerve is mounted in a plastic chamber so as to pass through five compartments (separated by grease-seals) containing oxygenated solutions. The first and fifth compartments are connected via calomel half-cells to a stable dc amplifier (chopper amplifier, time constant about 1 sec). The first compartment usually, and the second, fourth, and fifth compartments always contain a 5 mM K^+ Ringer's, in the experiments to be described. The nerve is stimulated before it enters the first compartment and the impulses are blocked in the third which contains a choline chloride (sodium-free) Ringer's. By this arrangement, it is possible to follow the changes in potential developed by an active or recovering region of nerve, the steady potential of an inactive region serving as a reference.

Figure 6 shows the potential changes observed during and following a 25-min period of activity at 50 volleys/sec. The downward deflection at the beginning of the tetanus is the time average of the negative action potentials. Once each minute the tetanus was interrupted for 5 sec and the recorded potential showed a positive deflection which reached its peak within this

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

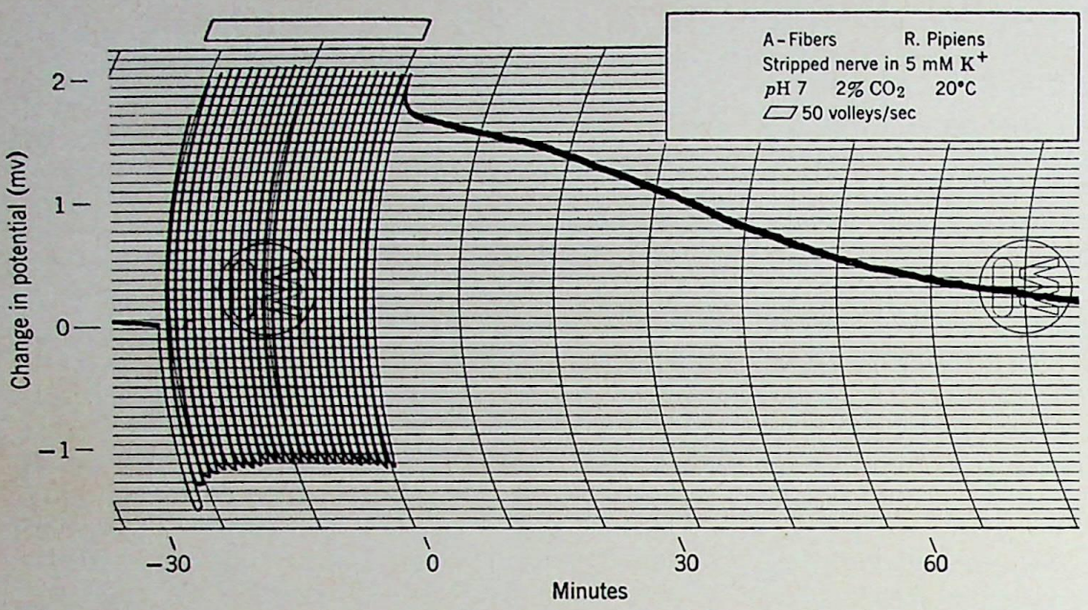


FIG. 6. Changes in average membrane potential recorded during and after 25-min tetanus. Tetanus interrupted for 5 sec each minute [from C. M. Connelly, "Post-tetanic hyperpolarization in frog nerve," in *Proceedings of the National Biophysics Conference, 1957* (Yale University Press, New Haven, to be published)].

interval. The amplitude of this hyperpolarization (net increase in membrane potential) reached a maximum after about 4 min and thereafter changed little. After the end of the tetanus, the hyperpolarization lasted for more than an hour; its recovery time course cannot be described as exponential.

Figure 7 shows how the prolonged positive after-potential is affected by the level of K^+ in the bathing solution (in the first compartment). A stimulation of 25 min in a solution containing 5 mM K^+ was first carried out as a control, and after recovery the solution was changed either to 8.5 mM K^+ (upper) or to K free (lower). The potential changes recorded during tetanus under the modified conditions have been superimposed on the control observations. The amplitude of the hyperpolarization varies appreciably with the level of potassium; recovery is rapid in high potassium and appears to be very slow and prolonged in K-free solution. During recovery from activity, the kinetics of oxidative recovery and the kinetics of the positive after-potential are both affected in much the same way.

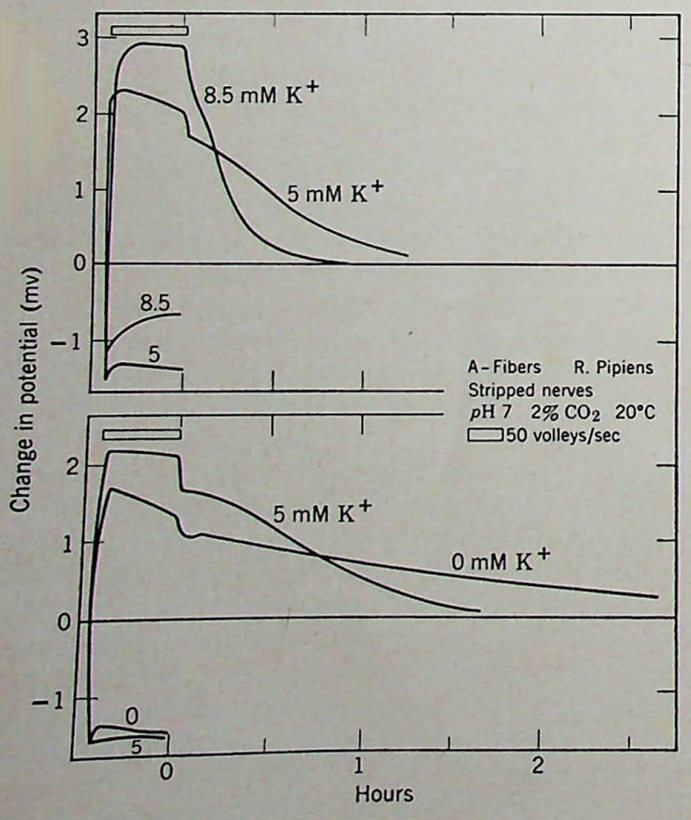


FIG. 7. Superimposed tracings of changes in average membrane potential recorded during and after 25-min tetani in Ringer's solutions containing different levels of K^+. Traces shown during period of tetanus are envelopes of changes similar to those shown in Fig. 6. In each of the two experiments the first tetanus was in Ringer's with 5 mM K^+, and the second tetanus carried out after 30 to 40 min exposure to the second solution. Zeroes of potential adjusted to superimpose at beginning of tetanus.

Other characteristics of this after-potential have been examined. It has been found that the observed amplitude of hyperpolarization approaches a maximum, or "saturation level" as the frequency of tetanus is increased above about 25/sec.[17] This maximum is about 2 mv. Depolarization produced by introducing isotonic potassium chloride into the first compartment averages 10 mv. Thus, maximum hyperpolarization corresponds to an increase in membrane potential of about 20%.

Post-tetanic recovery may be rapid or long-lasting depending on duration and frequency of tetanus. At frequencies above about 25/sec the effect of duration is striking, as illustrated in Fig. 8 which shows, superimposed, the recoveries from four tetani of different durations, at 50 volleys/sec. The longer the tetanus, the longer recovery takes, as if continued activity resulted in an accumulation of something, the final dissipation of which has associated with it an emf. Areas under the total hyperpolarization-time curves (i.e., area above the zero-potential line, including the periods of tetanus and recovery) have been measured as a function of duration of tetanus. Within experimental limits, area varies linearly with duration suggesting that the magnitude of hyperpolarization is approximately a linear measure of the rate of a recovery process.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

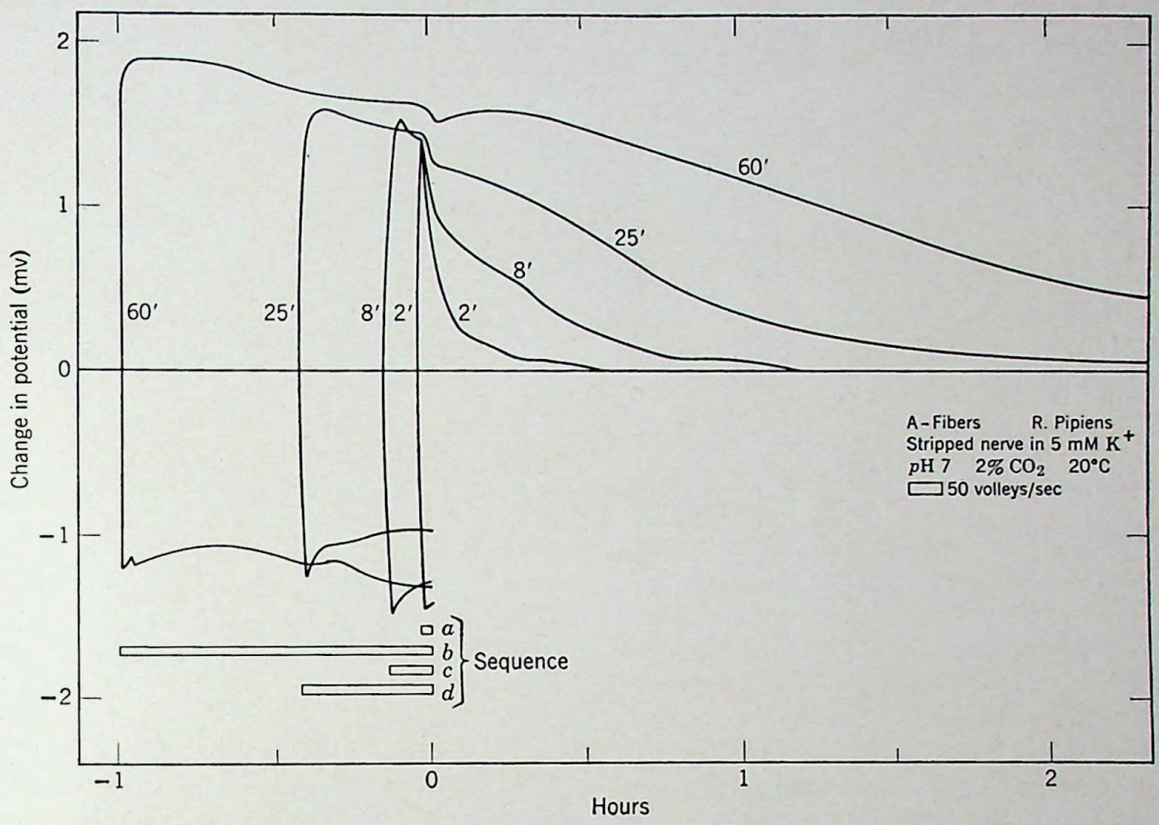


FIG. 8. Superimposed tracings of changes in average membrane potential recorded during and after a series of tetani of different durations. Traces shown during period of tetanus are envelopes of changes similar to those shown in Fig. 6 [from C. M. Connelly," Post-tetanic hyperpolarization in frog nerve," in *Proceedings of the National Biophysics Conference, 1957* (Yale University Press, New Haven, to be published)].

A variety of observations appears to be consistent with the idea that the recovery process involved is the outward transport or extrusion of sodium ions. One experiment which has a direct bearing on the question is illustrated in Fig. 9. The positive afterpotential associated with a tetanus is effectively eliminated upon complete substitution of lithium for sodium in the bathing Ringer's. The parallelism between this result and that described earlier on the effect of lithium Ringer's on activity respiration furnishes strong support for the idea that the activity respiration and the positive after-potential both have their origin in the process involving the outward transport of sodium coupled to the inward movement of potassium.

One further point should be discussed. Ritchie and Straub,[18] in studying the positive afterpotential of

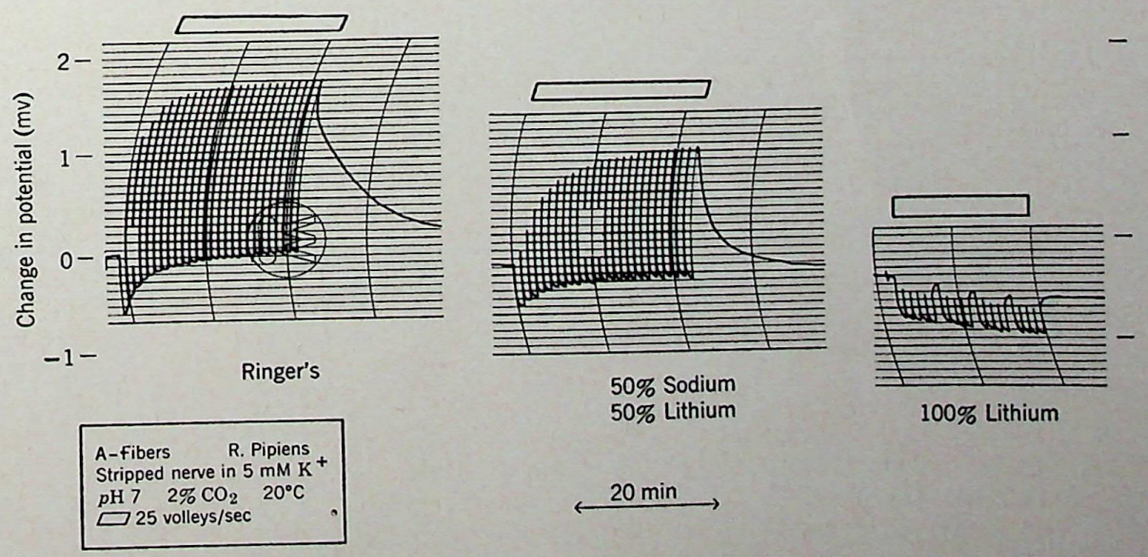


FIG. 9. Changes in average membrane potential recorded during and following tetani in solutions having different proportions of sodium and lithium ions. Potential zero common to the three records. During intervals (about 30 min) between records, nerve was bathed in the next solution. First and second records, 25-min tetani interrupted for 5 sec each minute. Third record, 20-min tetanus with 13 interruptions of 5 sec and 3 interruptions of 1 min each [from C. M. Connelly, "Post-tetanic hyperpolarization in frog nerve," in *Proceedings of the National Biophysics Conference, 1957* (Yale University Press, New Haven, to be published)].

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

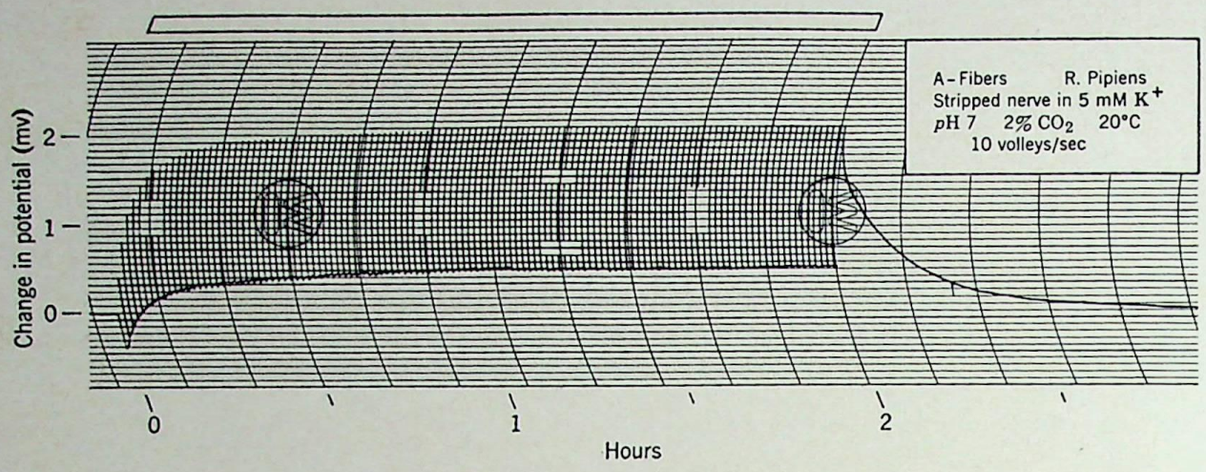


FIG. 10. Changes in average membrane potential recorded during and after 2-hr tetanus at 10 volleys/sec. Interrupted 5 sec each minute.

mammalian C-fibers, came to the conclusion that their observations could be explained solely by an electroneutral pump operating to restore the normal sodium-potassium balance following activity. The after-positivity was described as a variation of membrane potential in response to the variation in the concentration of potassium in the fluid immediately outside the nerve membrane. This happens as follows: during recovery, the pump produces a net flow of potassium inward across the membrane. By this action, the concentration of potassium at the outside surface of the fiber is lowered below its value in the body of the bathing solution. Since it is known that a decrease in external concentration of potassium does produce an increase in membrane potential, the conclusion appears to be unassailable, qualitatively. One must agree that any ion pumping system, whether electroneutral or inherently electrogenic, must have this property. It is to be taken

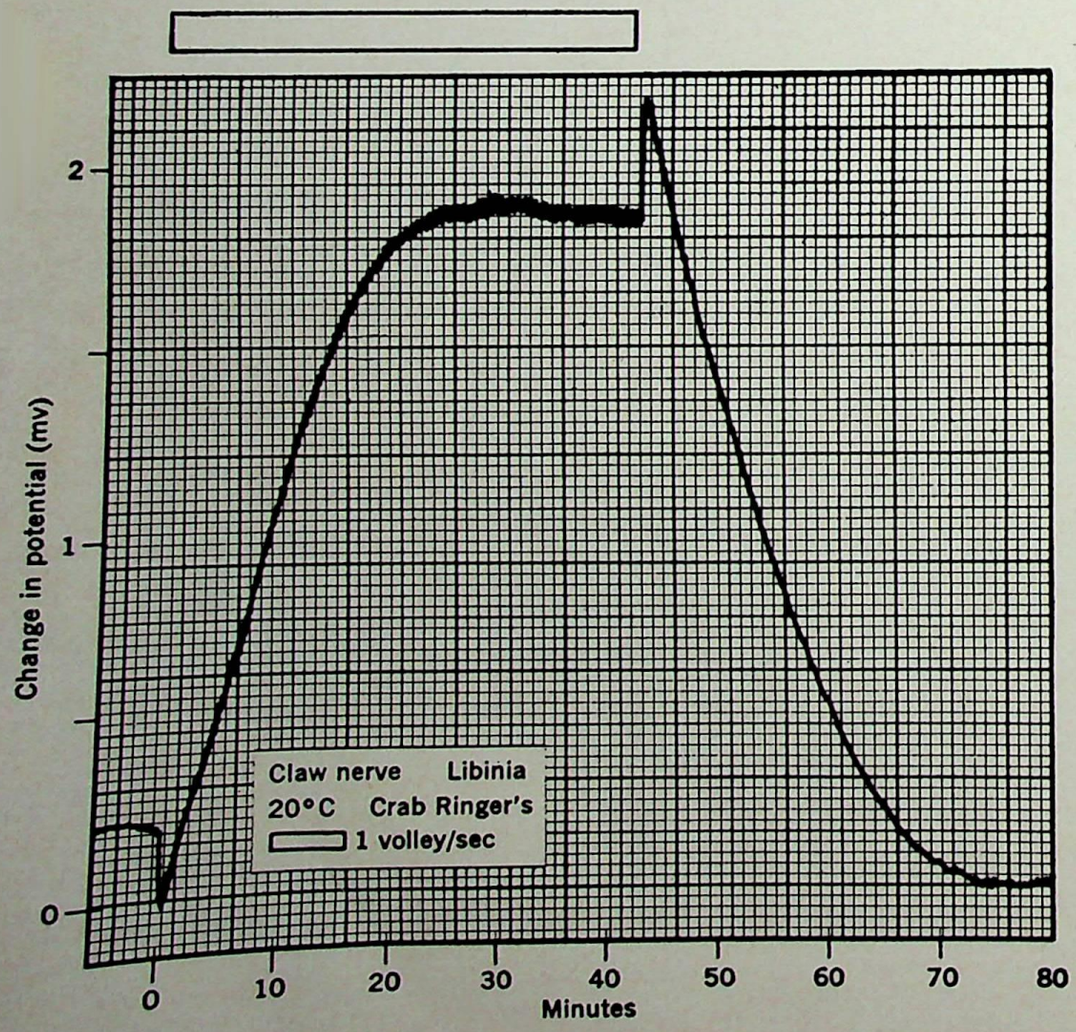


FIG. 11. Changes in average membrane potential of spider-crab nerve recorded during and following 43-min tetanus at 1 volley/sec. Not interrupted during tetanus. Measured depolarization of spider-crab nerve in isotonic potassium chloride about 30 mv.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

for granted that the transport system in frog nerve produces some potential by this mechanism. On the other hand, Fig. 10 illustrates an experiment in which the Ritchie and Straub mechanism is not sufficient to explain the observed variations of potential. During a tetanus, the average membrane potential (exclusive of the action potentials) should be either negative or zero, according to the Ritchie and Straub scheme. The average should be negative during periods in which there is a net loss of potassium from the fibers and would approach zero if the system approached a steady state (i.e., K^+ pumped back inside during pulse interval $\equiv K^+$ lost/impulse). The lower envelope of the record in Fig. 10 shows that the membrane became hyperpolarized four minutes after the beginning of the tetanus and remained so for the remainder of it. With correction for the apparent depolarization introduced by averaging the action potentials, the statement could be made that hyperpolarization began at the beginning of the tetanus and reached a final average level (in the 100-msec interval between pulses) of almost one millivolt. Thus, the ionic transport system of frog nerve appears to be inherently electrogenic (positive outward).

A similar statement, based on similar evidence applies to unmyelinated limb nerve of the spider crab, shown in Fig. 11. Here the membrane also develops a net hyperpolarization during low-frequency tetani.

If one takes as a working hypothesis the proposition that the magnitude of hyperpolarization is proportional to the rate of sodium extrusion, how can the shapes of the recovery curves of Fig. 8 be explained? The kinetic behavior of a simple reaction mechanism showing saturation kinetics is portrayed in Fig. 12. It is the classical enzyme-substrate reaction of Michaelis, the rate of which is given by the second expression in the figure, where K_M is the concentration of substrate that produces half-maximal rate. The curve describes the rate of reaction as a function of time as an initially large concentration of substrate decreases. If nerve were to extrude the sodium remaining within it with the kinetics of this mechanism, the hyperpolarization should follow this curve exactly, starting at a point appropriate to the

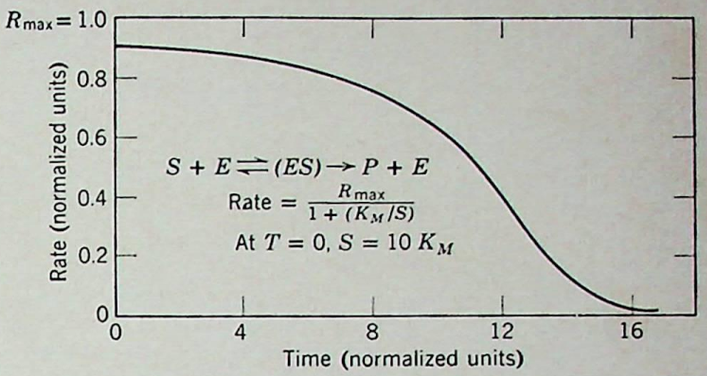


FIG. 12. Calculated time course of rate of disappearance of substrate with Michaelis type of enzyme-substrate reaction mechanism. K_M is concentration of substrate producing half-maximal rate [from C. M. Connelly, "Post-tetanic hyperpolarization in frog nerve," in *Proceedings of the National Biophysics Conference 1957* (Yale University Press, New Haven, to be published)].

amount of sodium accumulated during the tetanus. But, the recovery curves of Fig. 8 cannot be superimposed by displacement along the time axis, and, therefore, do not conform to this model.

If sodium ions enter the active fiber at the nodes and are also extruded principally from the nodes, then ions not extruded immediately after entry must tend to diffuse into the myelin-insulated internodal space. To determine the possible effect of diffusion in the internode on the kinetics of ionic recovery, the analog model shown in Fig. 13 was constructed. The model corresponds to one-half a node and to the adjacent one-half internodal space. A pentode, whose plate current-plate voltage curve is approximately a hyperbolic saturation curve of the Michaelis type, simulates the ion-extruding mechanism at the node; a filter network of 10 *RC* units is the analog of the diffusion field in the internode. Positive charges correspond to sodium ions; potentials at various points in the plate and filter circuits to sodium concentrations; plate current to rate of sodium extrusion; and injection of a constant current to the plate circuit simulates a tetanus. Seconds in the model correspond to minutes in the nerve. Figure 14 shows superimposed records of the output of the analog circuit, with parameters chosen to duplicate as closely as possible the

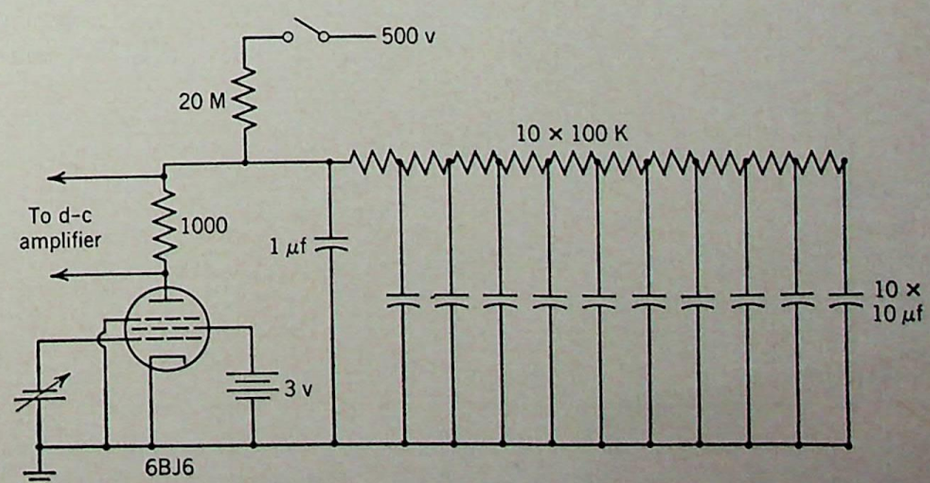


FIG. 13. Analog circuit [from C. M. Connelly, "Post-tetanic hyperpolarization in frog nerve," in *Proceedings of the National Biophysics Conference, 1957* (Yale University Press, New Haven, to be published)].

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

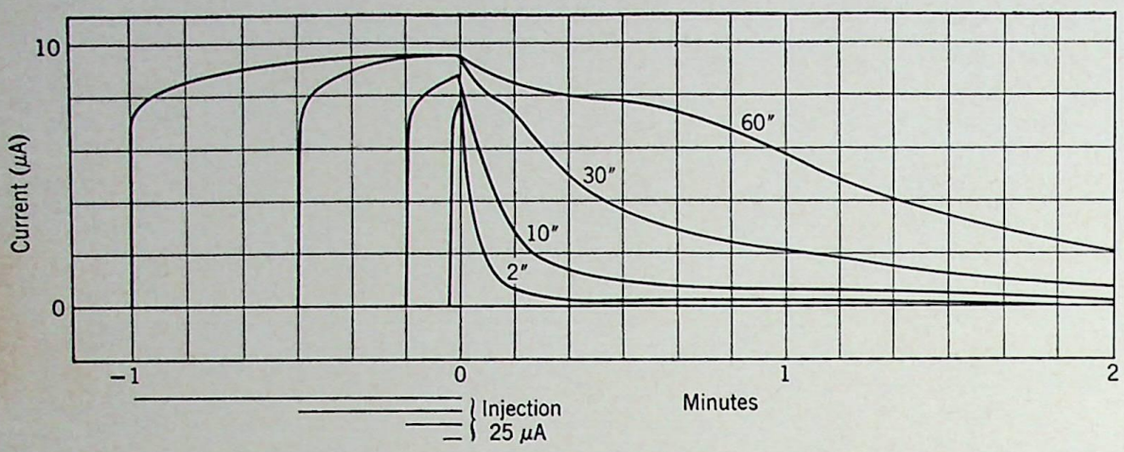


FIG. 14. Superimposed records of changes in plate current of analog circuit of Fig. 13, during and after 25-μA injections of different durations. Grid bias adjusted to give maximum plate current of 10 μA [from C. M. Connelly, "Post-tetanic hyperpolarization in frog nerve," in *Proceedings of the National Biophysics Conference, 1957* (Yale University Press, New Haven, to be published)].

observations of Fig. 8. The time constant of the network in this case is about twice as large as the corresponding time constant for the diffusion of sodium chloride in the internodal space (estimated assuming free aqueous diffusion and node-to-node spacing of 2 mm). It is encouraging that the shape of the recovery curve changes more or less in the proper manner as the duration of tetanus is increased. This oversimplified analysis tends to support the hypothesis that the positive after-potential has its origin in a saturable, ion extruding mechanism operating at the nodes.

CONCLUDING REMARKS

The events of recovery may be described tentatively s follows: Sodium ions enter a fiber during activity. The increase in internal concentration accelerates a saturable ion-transport process in which sodium ions and high-energy phosphate compounds are obligatory participants. Sodium ions are liberated to the exterior, and potassium in the exterior medium participates in the over-all cycle, being transferred to the interior. The operation of the mechanism generates an emf, directed positively outward. Oxidative metabolism maintains the supply of high-energy phosphate compounds.

The observations outlined in this discussion have emphasized the coupling between oxidative metabolism as an energy source and ionic transport processes as an energy sink. The molecular mechanisms of active ionic transport across membranes are almost a complete mystery. Proposed mechanisms range from shuttling carriers to micro-pinocytosis to enzymatic modification of pore configuration or charge distribution. This is an outstanding problem in cellular and molecular biology today.

BIBLIOGRAPHY

[1] F. Brink, in *Metabolism of the Nervous System* edited by D. Richter (Pergamon Press, New York, 1957), p. 187.
[2] P. D. Caldwell and R. D. Keynes, J. Physiol. **137**, 12*P* (1957).
[3] F. D. Carlson, F. Brink, and D. W. Bronk, Rev. Sci. Instr. **21**, 923 (1950).
[4] F. Brink, D. W. Bronk, F. D. Carlson, and C. M. Connelly, Cold Spring Harbor Symposia **17**, 53 (1952).
[5] C. M. Connelly, D. W. Bronk, and F. Brink, Rev. Sci. Instr. **24**, 683 (1953).
[6] T. P. Feng and A. V. Hill, Proc. Roy. Soc. (London) **B113**, 356 (1933).
[7] A. V. Hill, Proc. Roy. Soc. (London) **B113**, 345 (1933).
[8] B. C. Abbott, A. V. Hill, and J. V. Howarth, Proc. Roy. Soc. (London) **B148**, 149 (1958).
[9] A. L. Hodgkin and R. D. Keynes, J. Physiol. **128**, 28 (1955).
[10] A. L. Hodgkin, Proc. Roy. Soc. (London) **B148**, 1 (1957).
[11] I. Tasaki, *Nervous Transmission* (Charles C. Thomas, Springfield, Illinois, 1953).
[12] R. Stämpfli, Physiol. Rev. **34**, 101 (1954).
[13] R. C. Swan and R. D. Keynes, Abstr. Comm. 20th Int. Physiol. Congr. 869 (1956).
[14] W. P. Hurlbut, J. Gen. Physiol. **41**, 959 (1958).
[15] T. Asano and W. P. Hurlbut, J. Gen. Physiol. **41**, 1187 (1958).
[16] B. Chance and G. R. Williams, Advances in Enzymol. **17**, 65 (1956).
[17] C. M. Connelly, "Post-tetanic hyperpolarization in frog nerve," Proc. Natl. Biophys. Conf. (1957) (to be published).
[18] J. M. Ritchie and R. W. Straub, J. Physiol. **136**, 80 (1957).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

52
Sensory Performance of Organisms*

WALTER A. ROSENBLITH

Department of Electrical Engineering and Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge 39, Massachusetts

THIS paper is concerned with those aspects of organized neural complexity that permit organisms, such as man, to deal successfully with their sensory environment. Examples presented below illustrate how man, as an organism, detects, orders, and identifies events in his environment to which his sense organs are sensitive. The roles that stimulus intensity and time play in these operations are emphasized, as opposed to those aspects of sensory quality (pitch and hue, for example) for which one might expect the underlying mechanisms to differ rather considerably in going from one sensory modality to another.

Studies of communication processes in the nervous system should be based on a realistic view of the way in which the total organism reacts when it reaches selectively into its surroundings—be it under self-instruction or under instructions from others—to process stimuli that have informational value. This emphasis upon the organism's behavior in communication tasks leads to a preoccupation with certain dependent and independent variables. It leads to an inquiry into the "operating characteristics" of the organism in order to be able to specify, albeit in a statistical manner, its maximum sensitivity, its resolving power, its dynamic range, its characteristics in the frequency and time domains, its information-handling capacity, and so forth.

The data presented in this paper come mainly from *contemporary* psychophysics. Subjects are not asked to introspect, but, rather, to indicate by standardized motor responses (most often verbal responses) whether they judge a stimulus to be present or not, whether they judge two stimuli to differ, whether they can order a set of stimuli or identify members of the set.

Assume that one is dealing with well-instructed cooperative subjects who are skilled in the execution of the expected response and who are familiar with the set of stimuli that will be presented to them.†

Assume also that there is agreement on a program of stimulus presentation and on a method of response analysis. Thus, one must decide whether the admissable response categories shall be simply: "yes" and "no," "same" and "different," "more" or "less," or whether they should include the set of natural numbers. There must also be rules for dealing with false responses (false-alarm rate), and, finally, one must decide whether or not to quantify the temporal aspects of responding in addition to recording the mere emission of responses.

Such considerations may seem unnecessary details that should be left to methodologists, but unless there is a realistic understanding of the measurement and quantification problems in fields like psychophysics and neurophysiology, one can hardly evaluate the store of knowledge they have produced, or understand the relation of such knowledge to the knowledge that exists in the several areas of biophysics.

INFLUENCE OF INSTRUCTIONS UPON THRESHOLD OF DETECTION

The task of detecting the presence of a stimulus includes the classical absolute threshold as well as the generalized "masked" threshold—i.e., detection of the stimulus against a background of sensory stimulation. Such "background noise" refers necessarily to the sense modality under study. Even the best isolation booths do not reduce to zero all unwanted sensory influx other than that under the experimenter's control.

It has long been known that instructions significantly influence both the shape and the anchor points of "frequency-of-seeing" or "frequency-of-hearing" curves. A recent study by Smith and Wilson[2] has forcefully illustrated this dependence upon the judgmental criteria that are suggested to observers. Ten observers were asked to report the presence of a tone in noise. The three curves of Fig. 1 show how the identical group of observers performed this task under three different sets of instructions. Examples of the instructions that were designed to induce a "conservative" or a "liberal" attitude in the reporting of signals follow:

Conservative groups:

"Keep in mind that it is important for you to be sure you hear the tone. In many cases, there will be no tone, and you will be making a mistake if you indicate that you hear one. None of the tones is really easy to hear, but don't push the switch unless you are sure that you heard the tone. If you are in real doubt as to whether or not you heard a tone, assume that there was none."

* This work was supported in part by the U. S. Army (Signal Corps), the U. S. Air Force (Office of Scientific Research, Air Research and Development Command), and the U. S. Navy (Office of Naval Research).

† Experiments parallel to those considered here could be conducted with monkeys or even with rats or pigeons, if appropriate behavioral techniques were used. However, few experimenters[1] have been willing to train their animals thoroughly enough, or to reformulate the required response in a way that was compatible with the animal's behavioral repertory. The short-cut of employing highly motivated and intelligent subjects permits experimenters to bypass a lengthy learning process.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

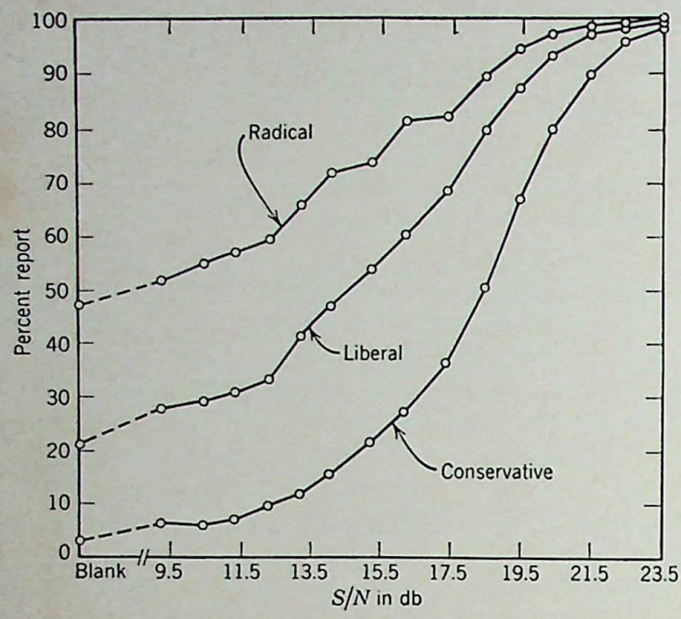


FIG. 1. Cumulative percentage of instances in which the presence of an 800-cps tone was reported as a function of signal/noise ratio. Attitude towards listening tasks is varied by means of instructions (see text). Each of 10 observers made approximately 150 judgments per point.[2]

Liberal groups:

"Keep in mind that all of these tones are hard to hear, and that you will rarely be absolutely sure that you heard something. But if you think you heard something, probably you did—report it. If you are very sure that you didn't hear anything—then don't touch the switch."

The results illustrate clearly that one cannot hope to predict accurately from a knowledge of signal-to-noise ratio alone the performance of a group of observers, even if the task calls for nothing more than the reporting of the presence of a single class of signals against the background of an unvarying noise.

Statistical models of the detection process have been formulated[3] in which performance is predicted not only on the basis of stimulus values and organismic variables; there are additional parameters that are sensitive to instructions and to amounts of reward or punishment. Such models may account for behavior such as is depicted in Fig. 1. A physiological interpretation of this sensitivity of the organism to nonstimulus variables remains to be given, though some physiological phenomena that provide grist for speculation have been reported in recent years.

When tasks that involve some ordering of stimuli are assigned to organisms, three problems are encountered. These stimulus-ordering operations may be listed under the headings of (a) differential sensitivity, (b) identification, and (c) psychophysical scaling.

DIFFERENTIAL SENSITIVITY

Such labels as just-noticeable difference (jnd), differential thresholds, and difference limens have been used to describe the changes in stimulus characteristics that make it possible for an observer to distinguish, with various degrees of certainty, between two stimuli. All of these measures of differential sensitivity are statistical in character and the jnd increases in size if a listener is required to identify correctly the more intense of two sounds in 90% of the trials than if he is required to be correct in 75% of the trials only.

More than a century ago, Weber formulated, on the basis of his own experiments, a generalization according to which the jnd between two stimuli gets larger as the stimuli get larger. This assertion has been called Weber's law, and the expression‡

$$\Delta I/I = k \qquad (1)$$

has been written to express this relative constancy of human differential sensitivity. Here, I is the intensity of the reference stimulus, ΔI represents the difference along the intensity dimension that is discriminable (in 75% of the trials, for example), and k is a constant. Weber's law states, therefore, that the jnd will remain proportional to the intensity of the reference stimulus. Today, there exist Weber functions (see Fig. 2) that cover a wide enough dynamic range to provide a fair test of the validity of Weber's law. The data indicate that, at least for auditory and visual stimuli, this relation breaks down as the absolute threshold is approached. Both frequency and intensity discrimination worsen as the stimulus becomes weaker. The existence of the so-called achromatic and atonal intervals[7] is further evidence along these lines. A better fit to the data of Fig. 2 is then provided by an equation of the form.

$$\Delta I = k(I + N_i); \qquad (2)$$

N_i has such a value that for low values of I, ΔI depends primarily upon N_i, while for medium and high values of I, N_i becomes negligible.

The introduction of N_i can be interpreted in different ways. Years ago, both Fechner and Helmholtz suggested that Weber's law be modified to acknowledge the presence of intrinsic interfering stimulation that could not be eliminated. In contemporary jargon, one might say that N_i corresponds to the "internal noise level" of the organism. This expression should, however, not be taken too literally. Several theoretical models exist in which a generalized noise term is a crucial feature of the discriminating organism. Thus far, modifications of Weber's law have aimed at agreement with the data near the absolute threshold. There is some evidence that the presence of extremely intense sensory stimulation leads again to some deterioration of discrimination. It is easy to see that the model could be further modified

‡ The symbolic notation employed in writing this Weber fraction or Weber ratio seems to emphasize the intensitive aspects of a stimulus. Actually, what is involved is not necessarily some aspect of energy flux, but rather loose usage of the term "intensity." Thus, fractions have been determined for lifted weights, length of line, pressure on the skin, and so forth, and comparisons between sense modalities have been made on the basis of these numbers.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

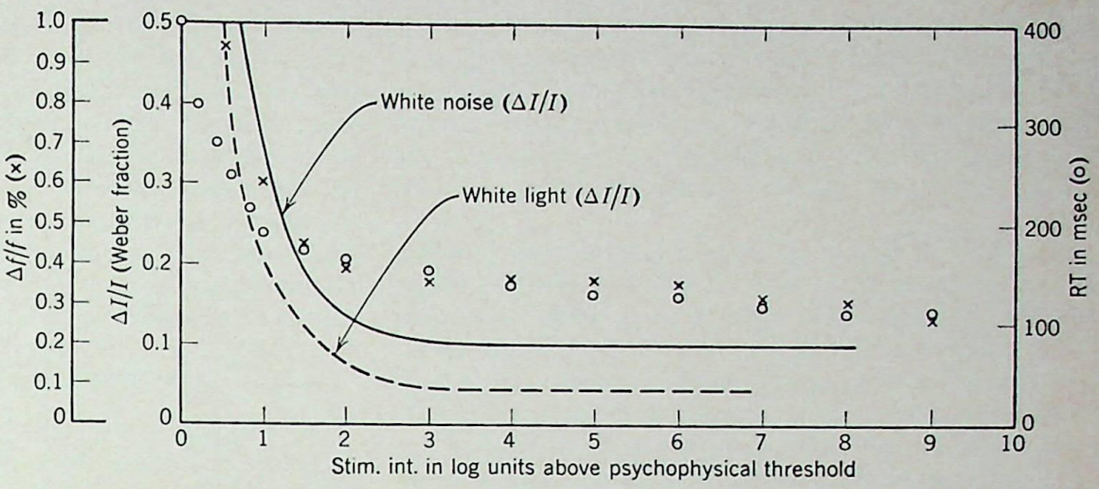


FIG. 2. Just-noticeable differences and reaction times as a function of stimulus intensity. Data on frequency discrimination of a 1000-cps tone after Shower and Biddulph[4]; on intensity discrimination for white light and white noise after Boring, Langfeld, and Weld[5]; data on reaction time after Chocholle.[6]

by the introduction of another term corresponding to an equivalent noise level that would only come into play for large values of I.

One further comment regarding the amount of invariance that it is realistic to expect in psychophysical data. Figure 3 summarizes the findings from more than half a dozen investigators, all of whom were trying to measure man's ability to discriminate pitch. In different experiments, different subjects, several psychophysical methods, several methods of stimulus presentation, and even several response criteria were employed. While it does not seem possible to read from this graph a value for the jnd for frequency, it is possible to point out that practically all of the data lie within the two dashed lines; i.e., their total spread is approximately a decade. The compilers of handbooks or of "Critical Tables on Sensory Performance" would thus be well advised to emphasize those aspects of psychophysical judgments that exhibit what might be called "relative invariance." Instead, they often insist upon quoting numerical values (sometimes with several significant figures), perhaps with the forlorn hope that the users of such tables will take these values with a grain of *ceteris paribus*.

SPAN OF ABSOLUTE JUDGMENT

From data on just-noticeable differences, estimates have been made of the total number of distinguishable sounds, lights, or smells. Usually, the authors of these estimates have simply divided the range over which an organism is sensitive by the corresponding average jnd. Thus, the "auditory area"[9] has been estimated to represent several hundred thousand distinguishable sounds, and the number of distinguishable colors has been estimated to be of the same order.

Several unjustified assumptions underlie these extrapolations. Subjects may find it quite easy to detect a difference between two stimuli without being able to identify either of them. In everyday situations, an individual rarely has difficulties in distinguishing between two other individuals even though he cannot tell which one is Jones and which one is Smith. Absolute identification requires much more information than differential discrimination.

Under the influence of the mathematical theory of communication,[10,11] with its concept of channel capacity, many experiments have been carried out to determine man's capacities to process selective information. By considering man as a communication channel with stimuli for inputs and responses for outputs, it has been possible to estimate maximal rates of transmission through him. The amount of input information can be varied either by varying the amount of information per stimulus or by varying the rate at which stimuli are presented. As Quastler[12] has pointed out, the maximal rate of transmission is limited by two factors: (a) people cannot emit more than 5 to 9 responses per second, and (b) people get confused when they have to discriminate

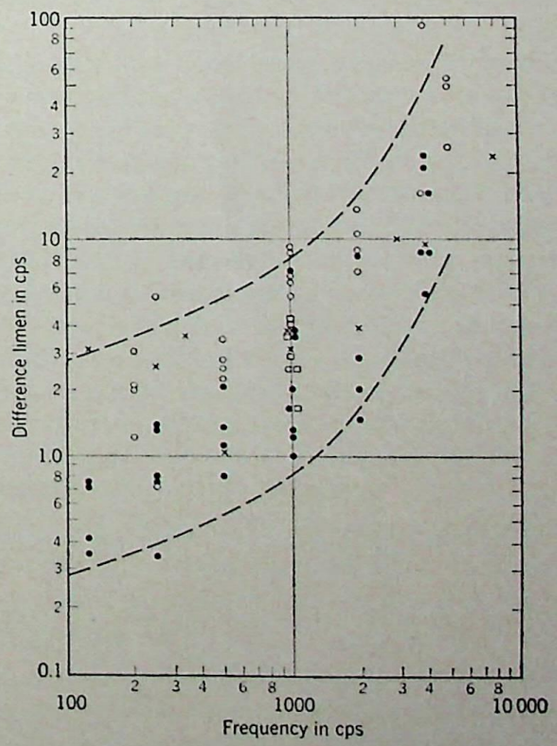


FIG. 3. Data on pitch discrimination reported by several investigators. Different symbols identify different experimental procedures (see Rosenblith and Stevens[8]).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

rapidly among too many alternatives. Thus, even under optimal conditions, transmission rates of more than 25 bits per second are hardly ever observed. The type of motor response required and the degree to which stimuli and responses are "compatible"[13] determine the particular figure that will be obtained for the transmission rate.

A different aspect of man's information handling capacity comes into focus if subjects are no longer required to respond as quickly as possible. They now need simply to tell which of several alternative stimuli occurred on a particular stimulus presentation. Given knowledge of the probability of occurrence for the different stimuli and of the accuracy with which the subjects perform, one can calculate the amount of information in these absolute judgments. This experimental paradigm has been much used in connection with "simple" stimuli, such as pure tones or monochromatic lights. The results from these experiments in several sense modalities have led to the emergence of what Miller[14] has called "the magical number 7 plus or minus 2." This number represents the number of alternative stimuli that subjects can accurately identify as long as only one aspect ("dimension") of the stimulus is varied. This magical number thus corresponds more or less to the upper and lower bounds for the "span of absolute judgment." More-precise estimates of performance figures will need to take account—in addition to individual differences—of such experimental conditions as the range of stimulus variation, the spacing of stimuli over a given range, and the number of available response categories. Subjects seem also to perform better if knowledge of results is fed back to them after each response.

Sensory communication in everyday situations is, however, not restricted to making unidimensional judgments. The sounds, shapes, tastes, and smells upon whose successful recognition one's sensory commerce with the environment depends are basically "multidimensional displays." Laboratory experiments with such displays indicate that the span of absolute judgments increases in size as an increasing number of stimulus aspects are varied. Pollack and Ficks[15] obtained 7.2 bits per judgment when they presented to their subjects sounds that varied along six different acoustic "dimensions." This performance corresponds to correct identification of approximately 150 different categories. Comparison of the Pollack-Ficks data (1.2 bits/judgment/dimension) with the characteristic figure for unidimensional judgments (2.3 bits/judgment) shows that, while each new dimension helps to transmit more information, these increments tend to diminish as the dimensionality of the display increases. But it is not known where the asymptote lies for this relation between bits per judgment and stimulus dimensionality. Thus, organisms seem much better in absorbing information from the environment by making several crude distinctions simultaneously, instead of by making extremely precise discriminations of a single aspect of a sensory stimulus. This finding will not surprise those who are familiar with the facts of speech communication where the precise spectral composition of speech sounds is rather unimportant in conveying information. Instead, man's ability to recognize speech sounds seems to depend upon the detection of the presence or absence of "distinctive features"[16] that are thus rather analogous to the dimensions of the acoustic display.

PSYCHOPHYSICAL SCALING

Among all the attempts to establish orderly relations between verbal responses and sensory stimuli, Fechner's law is probably the best known. By the use of indirect methods in which he used the just-noticeable difference as his scale unit, Fechner concluded that psychological magnitude varied as the logarithm of stimulus magnitude. There has been much argument about the measurability of sensation, about the validity of Weber's law, about the legitimacy of the Fechnerian integration of Weber's law which Fechner used to derive his own law; and yet, by and large, Fechner has been correct in predicting that his psychophysical edifice would stand because his detractors would not be able to agree on how to tear it down. In recent years, S. S. Stevens,[17] whose concern with measurement and scaling had led him to investigate systematically many of the fundamental assumptions in psychophysics, has produced a series of painstaking and ingenious studies of how subjects order stimuli by *direct* methods. Stevens did not accept the assumption according to which the unit of resolving power§ is the natural unit for experiments in which observers are asked to order a stimulus continuum by assigning numbers to it and in which these numbers are to reflect the judged or subjective magnitude of the stimuli. Stevens[18] divides perceptual continua into two general classes: continua that have to do with *how much* are called prothetic; continua that deal with *what kind* and *where* (position) are called metathetic. Well-known examples of prothetic continua are loudness and brightness; pitch is an example of a metathetic continuum. On the basis of experiments involving more than a dozen prothetic continua, Stevens concludes: ". . . there is a general psychophysical law relating subjective magnitude to stimulus magnitude . . . this law is simply that equal stimulus ratios produce equal subjective ratios. On numerous perceptual continua, direct assessments of subjective magnitude seem to bear an orderly relation to the magnitude of the stimulus. To a fair first-order approximation the ratio scales constructed by "direct" methods (as opposed

§ This indirect derivation of the scale of psychological magnitude parallels in some sense the extrapolated number of distinguishable sounds or colors. In both instances, the just-noticeable difference is the key concept of quasi-theoretical deductions without regard for operational compatibility.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

to the indirect procedures of Fechner) are related to the stimulus by a power function of one degree or another."[18]

Figures 4 and 5[19,20] present two samples of the numerous results obtained. In Fig. 4 there are plotted the magnitude estimates of loudness made by listeners who had not been given standard sounds of fixed intensity for comparison purposes. The eight different intensities were presented irregularly, and the subject estimated the loudness by numbers of his own choosing. These numbers were multiplied by appropriate factors to make each subject's estimate at 80 db the same, namely, 10. The points represent the medians of these "normalized" judgments, and the vertical lines represent the interquartile range of the judgments. By plotting the estimated magnitudes on a logarithmic scale *versus* the logarithm of stimulus intensity, the slope of the straight line is the exponent of Stevens' power function

$$\Psi = kS^n. \qquad (3)$$

Here, Ψ is the psychological magnitude, S is the stimulus magnitude, and k and n are constants. For the stimuli that have been studied, the approximate value of n varies from 0.3 for loudness and brightness‖ to 4 to 5 for electric shocks. The exponent for vibratory stimulation applied to the finger has a value of approximately 1. It is interesting to speculate that these exponents stand in inverse order to the dynamic ranges for the relevant sense modalities. The dynamic range for hearing and vision spans approximately 12 logarithmic units, the range for tactile vibration spans perhaps 3 to 4 logarithmic units, and the range for electric shocks spans only about 1 logarithmic unit. Does the fact that

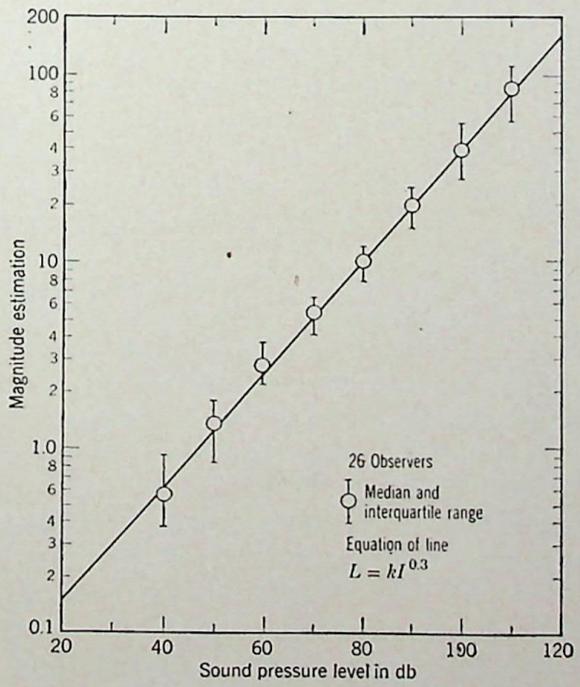


FIG. 4. Magnitude estimates of loudness made with no fixed standard.[19]

‖ The exponent is 0.6 if stimulus magnitude is expressed in terms of sound pressure instead of sound energy.

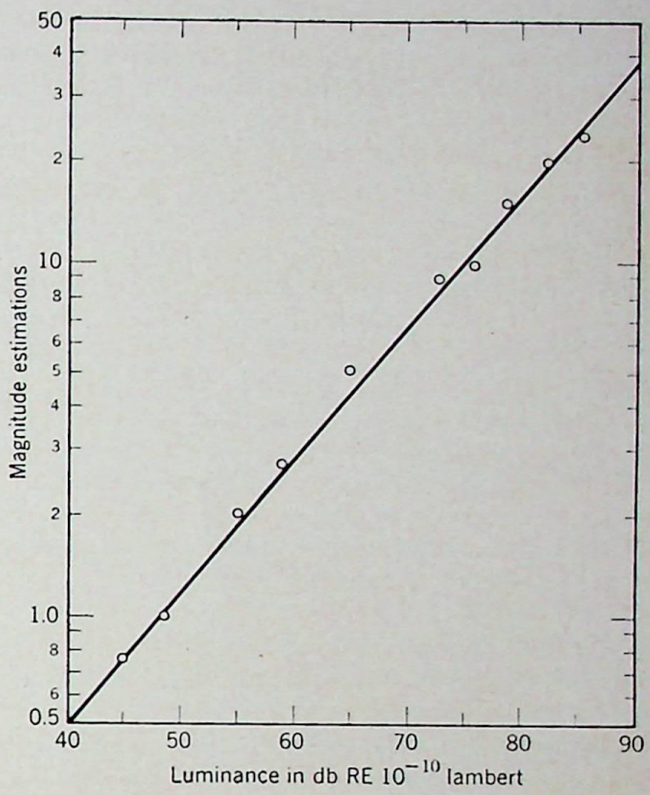


FIG. 5. Magnitude estimates of brightness of a luminous spot; the median estimates of ten observers are plotted.[20]

the product of the exponent multiplied by the dynamic range yields numbers of the same order of magnitude have any deeper significance? Does it reveal anything about the functioning of the sense organ, about the functioning of the nervous system, or does it rather reveal a certain constancy in verbal behavior—i.e., about the way in which people use numbers between the threshold of detection and the threshold of pain?

Another recent study by Stevens[21] deserves to be reported in this context. Once scales of subjective intensity have been established for several sensory continua, can these scales be used to predict what people will do when asked to match directly the loudness of a noise against the subjective intensity of a vibratory stimulus or an electric shock to the finger? Cross-modality matches were made between each modality and each of the other two. To a good approximation, the predicted functions were confirmed by direct experimentation, and the internal consistency of the direct approach to scaling has been further validated by this cross-modality study.

TEMPORAL CHARACTERISTICS OF SENSORY PERFORMANCE

Since Helmholtz's time, experiments on reaction time have been used to make inferences on the processing of sensory signals in the nervous system and on the duration of so-called mental processes such as discrimination, recognition, and choice. It was established that reaction times to sounds, electric shock, and touch were faster than reaction times to visual stimuli; also, that reaction

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

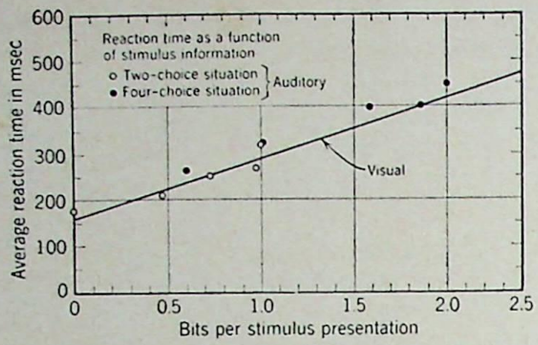


FIG. 6. The straight line represents the best fit to data obtained by Hyman[23] from one of his subjects in response to visual stimuli; the open and closed circles represent reaction-time data obtained by Albert[25] from one of his subjects in response to auditory stimuli; here, the stimulus set was composed of either two or four elements.

times decrease—at first rapidly and then more slowly—as the intensity of a sensory stimulus is raised above threshold (Fig. 2).

In addition to these data on simple reaction times, other experiments have attempted to assess the effect of task complexity on the speed of more discriminative reactions. It has been shown that reaction time increases as the stimulus ensemble increases in size and as its members become harder to distinguish from each other. This whole topic had lain more or less dormant for several decades until it became obvious that the application of certain concepts from information theory might make it possible to explore another aspect of man's capacity to handle information. A series of experiments[22–24] indicated that, at least over a certain range, choice time (or disjunctive reaction time) was proportional to the average amount of information transmitted per stimulus-response event. This finding applies equally well to visual as to auditory stimuli (see Fig. 6), and there is sufficient agreement between the data obtained by different experimenters that it is possible to talk about a characteristic figure for the "rate of gain of information" whose value lies near 6 bits per second. It is not possible to account for these time intervals in terms of delays on either the afferent or the motor sides of the nervous system. Enough is known about the latencies of evoked electric activity to permit one to say that these events consume only a fraction of the time involved in even the fastest reaction. The discrepancy becomes much greater, of course, upon considering sensory stimuli that are barely above the absolute or masked threshold or reaction times that involve complex discrimination. Thus, one must infer that a large fraction of the total reaction must be spent in "central information processing."

Recent experiments by Davis[26,27] and others throw some new light on these inferences. Davis presented to his observers two visual stimuli in succession and asked them to respond to each stimulus in turn. He then varied the interval between his two stimuli randomly between 50 and 500 msec. As long as the separation was more than 250 msec, the reaction time to either stimulus was approximately 200 msec; but when the time interval between the two stimuli fell below 250 msec, the second reaction time increased sharply. The second reaction time reaches its maximum value, approximately 400 msec, for an interval of 50 msec. The same phenomenon occurs when one stimulus is visual and the other is auditory, no matter what their order. Next, the subjects were instructed not to respond to the first of two lights. The results show little improvement over the delays that occurred when motor responses were made to both signals; the only difference was a slight shortening of the second reaction time, by about 20 to 30 msec. In a final experiment, the subjects were asked to make a "response" without a signal having occurred. Following the pressing of the key, a light came on (after an interval that was again randomly varied between 50 and 500 msec) to which the subjects were required to respond. This time there was no additional delay for any time interval. The subject was just as efficient in responding to a signal that came 50 msec after he pressed the first key as he was in dealing with a signal that came 500 msec afterwards. It would thus appear that the actual performance of the response does not contribute much to the delays that are observed, but that these delays are imputable to the processing of a signal that is still going on when the second signal is presented. There is further evidence that emphasizes the amount of time that central processes play in the handling of sensory information. In hearing and in vision, there is almost perfect integration of energy at the absolute threshold, as long as the stimulus is being presented for time intervals that are shorter than 200 msec. There is a whole class of phenomena in both hearing and vision in which a strong succeeding stimulus interferes with the "proper" analysis by the nervous system of a preceding weaker stimulus. The shorter the time interval between the two stimuli and the greater the discrepancy in intensity, the more severe is the interference. Since the delay times on the afferent side of the nervous system are longer for weak stimuli than for strong ones, the more intense stimulus has a chance to "catch up." However, the time intervals involved are sufficiently long that these differences in delay times along the ascending pathway are hardly the whole story. There is the further fact that interference with the more complex discrimination seems to be effective over longer time intervals than with more primitive tasks (such as simple detection of a preceding stimulus).

The preceding illustrations all point to the conclusion that the analysis of sensory information is a time-consuming task that occupies an organism's central nervous system in relation to the amount of information to be processed. This is hardly surprising; yet, it is often overlooked in practical situations.

The foregoing briefly reviews man's sensory performance as it appears when examined by the quantitative methods of contemporary psychophysics. Organisms

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

are able to deal with the sensory bombardment that impinges upon them by being satisfied to make crude discriminations, unless specifically instructed to make precise ones (at the cost of spending more time making them and of neglecting other sensory events than those under analysis). The flexibility of organisms in switching from one sense modality to another, from one aspect of sensory display to another, is bought at the expense of relatively low rates for the transmission of information. The mere existence of organisms seems to entail a high, internal noise level that prevents responses to environmental stimuli from being anything but statistical in character. As long as thresholds of tolerance are not approached, the more energy a sensory stimulus delivers to an organism, the higher is the probability that the signal will be processed in the central nervous system with both discrimination and dispatch. One may paraphrase this situation by saying that the higher the energy level of a sensory signal, the greater the redundancy with which the representation of this signal will be handled in the central nervous system. These are some of the considerations that led us to select certain aspects of the electrical activity of the nervous system for quantitative study.

BIBLIOGRAPHY

[1] D. Blough, J. Comp. and Physiol. Psychol. **49**, 425 (1956).

[2] M. H. Smith and E. Wilson, Psychol. Monogr. Gen. and Appl. **67**(9), 1 (1953).

[3] W. P. Tanner and T. G. Birdsall, J. Acoust. Soc. Am. **30**, 922 (1958).

[4] E. G. Shower and R. Biddulph, J. Acoust. Soc. Am. **3**, 275 (1931).

[5] E. G. Boring, H. S. Langfeld, and H. P. Weld, *Foundations of Psychology* (John Wiley and Sons, Inc., New York, 1948).

[6] R. Chocholle, Ann. oto-laryngol. **71**, 379 (1954).

[7] I. Pollack, J. Acoust. Soc. Am. **20**, 146 (1948).

[8] W. A. Rosenblith and K. N. Stevens, J. Acoust. Soc. Am. **25**, 980 (1953).

[9] S. S. Stevens and H. Davis, *Hearing: Its Psychology and Physiology* (John Wiley and Sons, Inc., New York, 1938).

[10] C. E. Shannon and W. Weaver, *The Mathematical Theory of Communication* (University of Illinois Press, Urbana, 1949).

[11] G. A. Miller, Am. Psychol. **8**, 3 (1953).

[12] H. Quastler, editor, *Information Theory in Psychology* (Glencoe Free Press, Urbana, Illinois, 1955).

[13] P. M. Fitts and R. L. Deininger, J. Exptl. Psychol. **48**, 483 (1954).

[14] G. A. Miller, Psychol. Rev. **63**, 81 (1956).

[15] I. Pollack and L. Ficks, J. Acoust. Soc. Am. **26**, 155 (1954).

[16] R. Jakobson and M. Halle, *Fundamentals of Language* (Moulton Company, The Hague, 1956).

[17] S. S. Stevens, Science **127**, 383 (1958).

[18] S. S. Stevens, Psychol. Rev. **64**, 153 (1957).

[19] S. S. Stevens, J. Acoust. Soc. Am. **27**, 815 (1955).

[20] S. S. Stevens and E. H. Galanter, J. Exptl. Psychol. **54**, 377 (1957).

[21] S. S. Stevens, "Cross-modality validation of subjective scales for loudness, vibration, and electric shock" (J. Exptl. Psychol., to be published).

[22] W. E. Hick, Quart. J. Exptl. Psychol. **4**, 11 (1952).

[23] R. Hyman, J. Exptl. Psychol. **45**, 188 (1953).

[24] E. R. F. W. Crossman and J. Szafran, Experientia Suppl. IV, 128 (1956).

[25] A. E. Albert, *Analysis of variance in reaction time experiments.* B. S. Thesis, Department of Mathematics, Massachusetts Institute of Technology, June, 1956 (unpublished).

[26] R. Davis, Quart. J. Exptl. Psychol. **8**, 24 (1956).

[27] R. Davis, Quart. J. Exptl. Psychol. **9**, 119 (1957).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

53
Biological Transducers and Coding*

OTTO H. SCHMITT

Departments of Zoology and Physics (Biophysics), University of Minnesota, Minneapolis 14, Minnesota

IN this discussion of biological transducers and coding, no attempt is made to describe in detail the specific codes in which diverse animal (and plant) transducers operate, for this attractive topic in comparative physiology and biophysics is discussed at length in other sections of this book by Bullock, Hartline, Katz, and others. Nor is an attempt made here to correlate over-all organismic behavior with transducer functions for that is, to a large extent, the burden of the communication by Rosenblith. Instead, the niche between organismic behavior and detailed transducer codings is examined to determine whether or not it is possible to develop new and useful analytical tools by which one can better formalize the behavior of individual transducer elements and their interrelationships, so as to approach a deeper and more intuitive understanding of the resulting whole organism's stability, decision making, intelligent behavior, and homeostatic performance.

While it is impossible to develop the point of view adequately in a few pages of text, it is the firm conviction of the author that biophysical theory, where it has developed at all, has developed too closely in the footsteps of physical theory and has not yet evolved mathematical models and analytical techniques specifically tailored to fit biological phenomenology. Instead of adapting physical-science terminology and theoretical formulations to its needs, it has adopted them as though they were "true" rather than merely convenient approximate descriptions of experimental experience. This has been especially true in the studies of biological transductive processes where the biologist, armed only with experimental facts, has for the first time come up against the new and highly formalized jargon of information theory and automatic-control engineering. Through modesty or lack of mathematical dexterity, he usually tries to make-do with these mental tools as furnished him ready-made by physics and engineering. It is not the author's purpose to deny that one must start with available techniques, but rather to propose that one must quickly amend and supplement these with uniquely biophysical hypotheses and an appropriate calculus, if the problems of this field are to be reduced to a complexity where intuitive insight into them can readily be gained.

* Much of the original work reported in this section was supported by grants H513 from the National Institutes of Health, Public Health Service, U. S. Department of Health, Education, and Welfare, by contract N6onr-24609 with the Office of Naval Research, U. S. Department of the Navy, and by grants from the Minnesota Heart Association and the University of Minnesota Graduate School.

It is necessary, however, to set up some ground rules before launching into a discussion of biological transducers, their transfer functions and other possible means for studying their organization and coding. As there are no uniformly adopted definitions for transducers in biology and biophysics, an explanation is probably necessary for the term biological transducers. The usage of the author may be somewhat different from that of others in the biophysical field, and perhaps different even from that of other authors contributing to this volume.

What, then, is a biological transducer? The term transducer is borrowed from the engineering sciences, where it is usually defined so as to include those devices which "lead across" or transform energy from one form into another, and thus translate signals from one energy modality—such as sound—into another—such as a varying electrical current. While the physical-science definitions do not usually specifically exclude gross-energy converters such as steam boilers, storage batteries, and electric light bulbs from being considered as transducers, it is tacitly assumed that only devices which serve a communication or signal-converting function will be called transducers. Thus, a motor turning the rudder or shifting the ailerons of an airplane in response to instructions from the autopilot is clearly regarded as a transducer; one driving a washing machine is not. Similarly, a resistance-wire strain gauge is considered a transducer in its function of converting mechanical position changes into electrical-current variations, but its incidental function of converting electrical power into heat is not usually thought of as transduction.

Biological transducers, then, should logically be the biological counterparts of the physical sciences' transducers and should comprise all of those biological units and systems which perform transformations of energy modality within the context of a communication or informational transformation function. Thus, biological transducers should include not only the familiar sensory transducers such as the rods and cones of the retina, the thermal, olfactory, and gustatory receptors, the vestibular and cochlear receptors, and the many proprioceptive organs, but they should also include numerous synaptic and related intermediary units involved in biological-data processing and storage, and the output or motor transducers as well. Nor is there any reason to exclude humorally mediated transducers. The concept of a transducer fits the function of a cortical neuron or of the motor end plate at least as well as it does that of the muscle-spindle strain gauge.

It is convenient to use the terms afferent and efferent

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

in reference to biological transducers to distinguish incoming from outgoing transduction functions, and it seems desirable to adopt directly from engineering the terms "passive" and "active" transducers, to distinguish between those where the energy of an incoming signal provides directly the output energy and those transducers, very common in biological systems, where the incoming energy triggers or controls the release of locally supplied energy into an output modality.

Transducers that operate in one direction only are called unilateral transducers, while those that operate in either direction are called bilateral. The quite special transducers that not only operate in either direction, but also obey the reciprocity theorem, are called reciprocal transducers.

The linear transducer is one where the output can be related through linear-differential or algebraic equations to the input function. It is a type of transducer not often found in biological nature, but a form to which most experimental transducer results are approximated, as its behavior is easily formulated and understood mathematically. The linear transducer does deserve special mention, for it is out of the notion of a linear transducer that the transducer transfer-function idea arose.

The transfer function of a linear transducer is usually expressed simply as the ratio of its output to its input signal with frequency as a parameter. This ratio is usually a complex number which may be dimensionless in the case of a transduction between similar modalities, or may carry within it the ratio of the dimensions in which output and input are expressed. It is valuable to remember that, around any properly described, closed, biological control loop, transfer functions must cancel dimensionally. Thus, a mathematical description of a biological loop system can be tested much as elementary physics equations are tested dimensionally by cancellation of dimensions and by reduction to mass, length, and time units.

In studying biological transducers, it is often desirable to regard the transfer function as a much more general operator than merely a linear complex ratio, for many of the merits of transfer-function analysis remain even where the transfer operator is a very elaborate time- and pattern-dependent code.

It is evident from these definitions and the associated discussion that the foregoing concept of biological transduction is very inclusive. The term "biological transducer," is used here to include all of those organs and processes wherein the purpose of the reaction involved is to convey, store, or process information about the environment and organism for its preservation and effective operation. Included, therefore, are all of those processes whereby information is transduced or led across some kind of a transformative exchange, with no restriction, for example, to sensory receptor organs. In this definition, genetic determination, many processes of differentiation and growth, as well as reflex and higher level abstraction mechanisms, become specifiable in transductive terms. It may be felt that this is too inclusive a definition, for it obviously includes a great deal of biochemistry, hormonal control, enzyme dynamics, growth processes, and central nervous whole-organism behavior control. It can be limited, however, and made meaningful by adding the loose criterion that has entered into some European engineering literature to distinguish between what are called large energy processes and small energy processes, and by recognizing that this distinction is one of convenience rather than one of fundamental importance.

Behind this categorization is the idea that, while every energy turnover process, whether power-plant generation of electricity or ATP utilization in muscle, involves a certain amount of direction, and while every informational biological activity, even so explicitly directive as the axonal propagation of an impulse, requires the expenditure of energy, one can usually choose one of these as being of dominant importance in the particular process in question.

One often finds in complex control systems that up to about 10 or 15% of the total energy consumed will go into processes that may legitimately be considered as primarily informational. This fits reasonably well with biological systems where, if one considers the human as an example, something like 10 to 15 w of power goes into the total nervous-system maintenance and operation, whereas the total average power turnover for a normal-size person of average activity is somewhat over 100 w. It is sound economical biological design that one should put perhaps 10 to 20% of the available metabolic energy to work for learning about the environment, deciding what to do about it, and issuing proper instructions to do something appropriate.

Now, to turn to a different aspect of biological transducers and coding, it is rather hard to break from tradition to frame variables conforming to codes in which biological processes are described by the organism, as against those terms in which one is used to describing similar phenomena in the physical sciences. In engineering analysis, one very firmly separates time as a variable apart from space coordinates, whereas one is forced to conclude that the organism uses in almost every transducer a conveniently mixed code and pays very little attention to whether it separates out temporal and spatial aspects of pattern.

Figure 1 illustrates one of the classical sensory-receptor code patterns (that of the carotid-sinus pressure receptor) to show that these receptors do not slavishly report on the value of a function being transduced or on the time derivative of this function exclusively. Instead, they answer the biologically important questions: "What is the present state of affairs?" "What is changing?" and "What has been the recent past experience?" These reports correspond loosely to engineering transducer language in which function, time derivative, and integral transducers are involved, but in

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

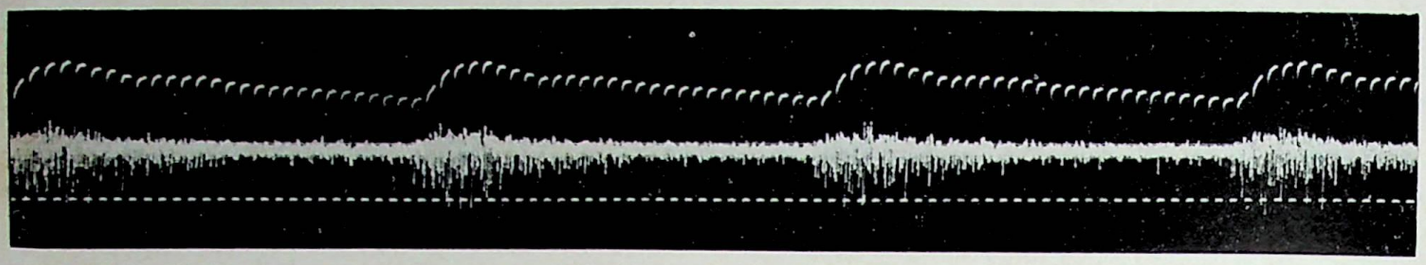


FIG. 1. Carotid sinus pressure transducer response (dog). Central record represents multifiber nerve response. Upper curve envelope is directly measured blood pressure referred to lower base line as zero. Mean blood pressure 150 mm Hg, interruption rate on pressure measuring trace 60/sec. Each pattern represents one heart beat (records by courtesy John W. Trank).

engineering transducers these factors usually are rather carefully separated out or very cautiously combined. Figure 1 is a record of the action potentials recorded from a few nerve fibers leading from the carotid-sinus end organs in a dog. This organ is concerned with reporting on the mean blood pressure in the animal and also on the details of pressure during the heartbeat. In this case, the average pressure is an important item and its absolute value remains important. Therefore, this must be a relatively nonaccommodating organ in the sense that it must continue to report the present state of affairs.

On a short time basis, it has a large "derivative" term, and thus is modulated from a high frequency almost to cutoff during each beat. Its steady-state response as reflected in the mean frequency, and especially in static records like those of Fig. 2 for a similar organ in a rabbit, is shown to convert static pressure almost linearly into pulse frequency. In this static sense, one might call this carotid-sinus organ a linear pulse frequency-modulation transducer. One must however, point out the danger of specifying a transducer as "linear" or "nonlinear" on this basis by referring to the two graphs accompanying Fig. 2 where data from the same impulse records are plotted in (a) as a frequency *vs* pressure-transfer relationship and in (b) as a pulse interval *vs* pressure graph. Note how the choice of variable determines "linearity" or "nonlinearity" in this casual sense. One must, therefore, define quite carefully on a theoretically sound basis what kind of nonlinearity is meant when that term is used.

This transducer an engineer might still consider reasonably well-behaved, but now consider the photoreceptors of Fig. 3 which use an amplitude, nonlinear heavily time-influenced code. These transducers operate on a pulse-code modulation characteristic much like that which would be anticipated for a leaky, resistance-condenser coupling decay characteristic. Frequency is very high initially and decays for constant illumination so that frequency increases, but not linearly, with intensity, and decays with time for any given intensity. In addition to this general pattern of response, some animal photoreceptors also give oppositely polarized responses so as to remark pointedly that the light is decreasing or going out; an event worthy of mention, even though in the negative sense. Some sensory transducers, such as those in the tongue concerned with measuring temperature, use their nerve frequency spectra doubly, once in an ascending sense, once in a descending, the range in use being determined by a separate signal.

It was only about fifteen years ago that it was discovered that "contrast" or accentuation of detailed changes in a one- or multi-dimensional field of transduced data could be greatly enhanced by systematic suppression in a readily specifiable manner by each region of every adjacent region. In fact, it was only in about 1945 that spatial, maximum, and minimum finding networks were first produced, temporal peak predictors having come only a little earlier.

Now it is found that a very similar system exists in optical, acoustic, and probably other sense-modality, biological transducer systems, so that much of the pattern finding is done before data reach the central nervous system.

Where there is ample signal energy to operate many transducers at substantial output, as in foveal bright-light vision, this detailed differential discrimination is common, while in threshold signal performance, adjacent units are often pooled to improve statistics. There is an interesting auditory case reported where noise heard in one ear can be used to subtract out coherent noise from noise plus signal in the other ear, to permit recognizing signal otherwise submerged below recognition. In this case, two ears cooperate in a field-subtractive type of selective filtering.

It is often stated that the sensory codes are essentially logarithmic in the sense that equal fractional increases in intensity of stimulation rather than absolute increases are required to evoke correspondingly detectable increases in response. While this generalization has been elevated to the status of a "law" and is often roughly true over a decade or two, it is usually noticeable that there is a threshold region where logarithmicity is not obeyed and that there is a bending over of even the logarithmically plotted line at extremely high intensities. In long-range senses, there will usually be one or more automatic-gain control systems operating to extend the working range. It will interest engineers and physicists to know that biological transducers often operate in pseudo-instability, much like that of the superregenerative radio receiver in the absence of signal, and share

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

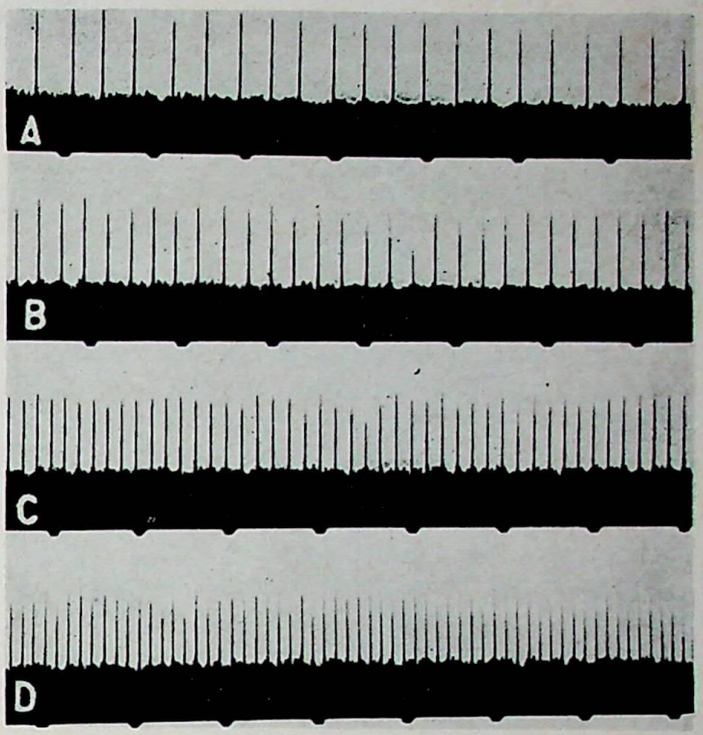
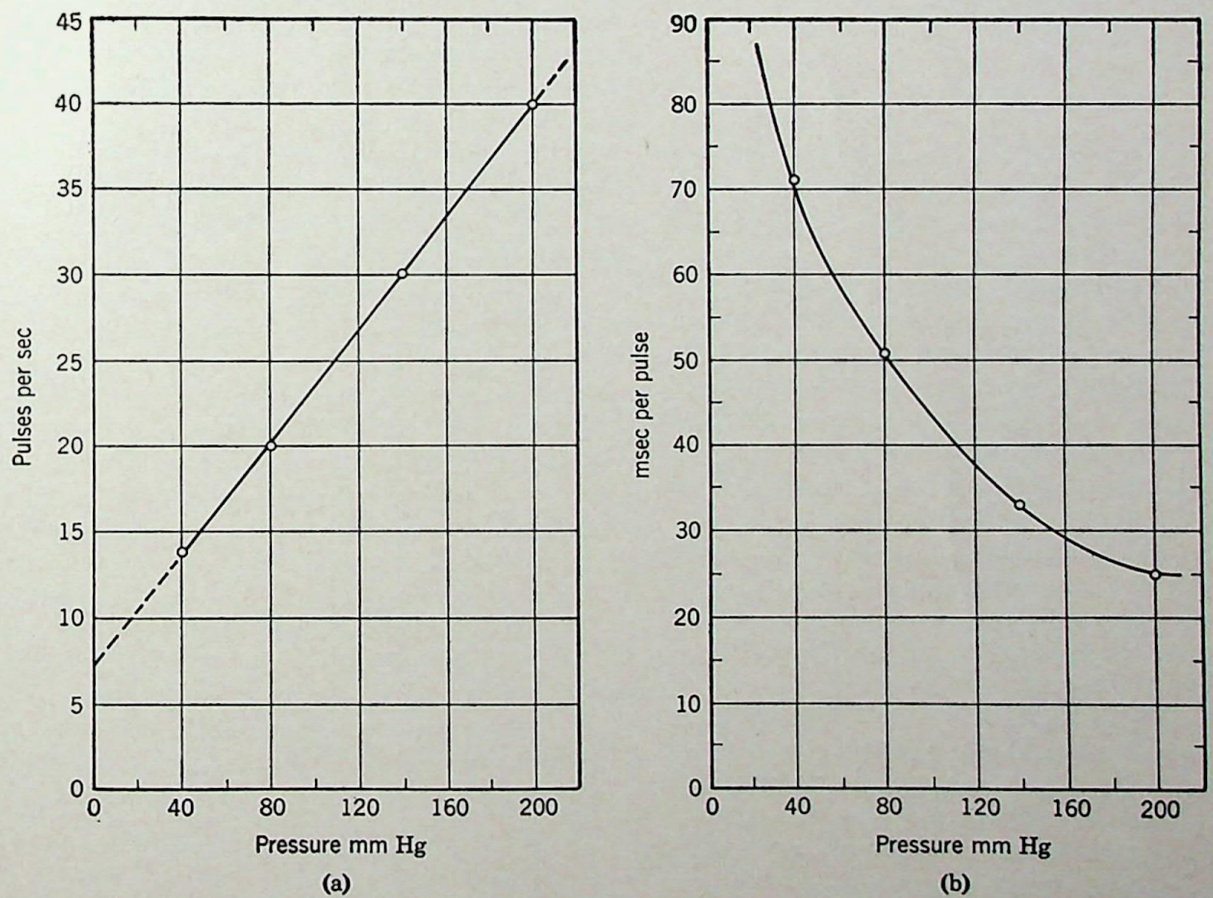


FIG. 2. Steady-state response of carotid sinus pressure transducer [from D. W. Bronk and G. Stella, Am. J. Physiol. **110**, 708 (1934)]. Data from original records of this figure are replotted in Fig. 2(a) as frequency *vs* pressure and in 2(b) as pulse interval *vs* pressure to emphazise importance of descriptive variable chosen in determining whether transducer is "linear" or "nonlinear."



with the superregen, exceedingly high sensitivity near threshold and a tendency to lock onto any furnished rhythms. This point again deserves attention in connection with biological rhythmic processes, mentioned briefly in a later section.

It is perhaps safer to generalize that the sense-organ receptors which report in neural pulse-code modulations usually spread the available frequency band width roughly equally, in terms of importance, over the continuum of intensity time space in which they have to report, than it is to assume that they adhere to the Weber-Fechner ideal of true logarithmic response. The accuracy with which a sense organ can report over a given neural channel depends on the length of time in

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

FIG. 3. Record showing adaptation of a sense organ (photoreceptor). Stimulus was maintained at constant strength throughout, but interval between discharges steadily diminished. Signal indicates duration of stimulus application and time is in 0.2-sec intervals [record of H. K. Hartline cited from D. W. Bronk, Research Publs., Assoc. Research Nervous Mental Disease 15, 60–82 (1934)].

which it is allowed to report for a given discriminating mechanism. In general, the organism must have its report quickly, to avoid vacillatory behavior and sluggishness, yet reasonably accurately over a wide range. Thus, the neural codes turn out, for the narrow band width in frequency of pulses available, to be apt compromises between adequate and quick reporting.

One is familiar with the interestingly interchangeable way in which temporal and dimensional variables enter into the ordinary wave equation describing the progression of a pattern or wave of form f in space x and time t. Equations (1) show this relationship for one dimension:

$$\begin{aligned} F &= f(x-ut) \\ \partial F/\partial x &= f'(x-ut) \\ \partial F/\partial t &= -uf'(x-ut) \\ \partial V/\partial t &= -u\partial V/\partial x, \end{aligned} \qquad (1)$$

but it is easy to see how it can be generalized for more dimensions. There is a linear relationship between the position of a particular feature of the wave in space and the passage of time.

As a practical illustration of this equivalence, it is recognized that one frequently interchanges mentally the picture of a nerve action potential in the space-time world. Usually, one measures the potential variations at one place along the nerve as a function of time, but, without feeling any strain, one substitutes the picture of a wave of identical form progressing smoothly down the length of the nerve as a fixed pattern.

It will certainly be important to discover the mechanisms within neural net systems whereby such interchanges become easy and in which this kind of spatial-temporal pattern analysis can be accomplished with the available neuronal networks and fields. It must be expected in this seemingly messy and complicated situation that the usual biological optimization of representation will apply. One must expect near-logarithmic compression of past and future in this representation, and similarly one must expect that spatial features will be compressed in their representation as they become farther removed pattern-wise.

It is attractive to think of the neural net as a multi-dimensional analytical space in which distance along partially excited, short neuron extensions corresponds to temporal or spatial separation. Figure 4 shows how the phase velocity of an unpropagated nerve impulse goes on at near-normal speed in a passive axon, ready at any time to be reinforced into partial or full excitation by surrounding fields. A corresponding lower-velocity process easily can be anticipated for the short neuronal extensions in the central nervous system. In this connection, it is worth mentioning that application of readily available, distributed negative-resistance networks to *in vitro* nerve preparations allows exploration of this type of reinforcement.

Consider now some basic differences in approach to informational coding that must apply in biological studies as contrasted to the conventional engineering analysis of comparable systems. In an engineering system, the coding is generally superimposed from without by the machine design, so that there is real justification for dissociating the message from the code. In biological systems, the organism must create the complex of a code plus the messages in this code with which it will function.

It consequently becomes possible in the biological case to compare meaningfully the cost of a very elaborate code, in which a one- or two-bit message can tell one to record the complete contents of the *International Critical Tables* or of the *Encyclopedia Britannica*, with a very elaborate message in a simple code whereby one dictates one or the other of these works at length. In a genetic plus sensory-motor biological code complex, the organism seeks to minimize the combination of these two aspects.

Where there is to be very frequent repetition of similar messages, it becomes economical to devise fairly elaborate codes in which these messages can be shortened and sharpened in discrimination. For the very rare or

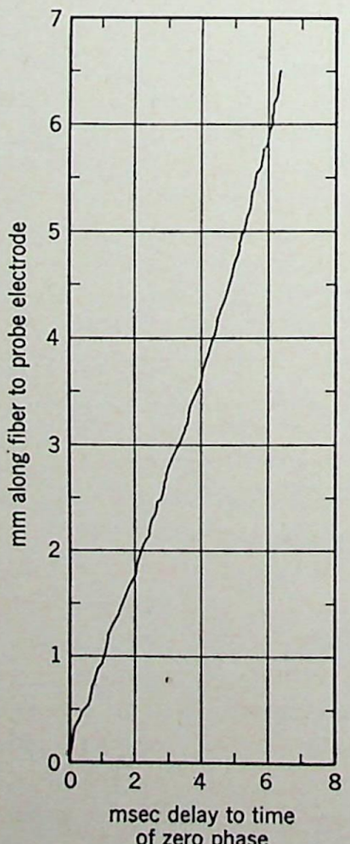


FIG. 4. Phase velocity of an unpropagated subthreshold sinusoidal wave in living axon. Data for squid axon, mean diameter 0.41 mm, 300 cps, 21°C. Phase velocity 10 meters/sec varying slightly with taper of fiber. Record is directly traced from machine drawn curve without smoothing.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

one-time situations, it is economical to spell out the message in detailed, simple versatile code so long as it is not essential that the communication be completed in a very short time. Undoubtedly, some of the complex innate behavioral responses of lower and higher organisms have to be spelled out rather expensively in informational, genetic, and developmental symbols, as they are essential to survival and can be left only slightly sketchy.

It is an almost universal observation that, wherever a specific biological organic structure or organization functionally need not be built to a firm pattern, it will be left to grow as dictated by local circumstances. This relieves the informational system of that much superfluous burden. If it is necessary only that there be a substantial number of muscle fibers or connective-tissue units in a particular structure, it is very improbable that an exact number will be present, while it is equally certain that a nine-plus-two or a correspondingly important topological or chemically significant structure will be firmly designed-in. The informational design, once having been paid for genetically, will be used widely, thus minimizing the necessity for superfluous duplicate instruction. Biological organisms are probabilistic mechanisms, and their transducer codings reflect their willingness to take a statistical chance for survival in a risky world, especially if the odds are good. Where an item of information is especially important, it will usually be protected by replicate and nonreplicate redundancy. Such redundancy may become too expensive in some cases and may have to be sacrificed. This may be the case in some genetic transformations where cells and indeed the continuity of life for the organism become susceptible, briefly, to minor external mutational influences.

To avoid further digression from specific, experimentally accessible material, some of this free-wheeling thinking should be tied down to mechanisms and bits of analysis where there is a chance for specific solutions to specific problems. This set of analytical tools is still rather dull, but undoubtedly can be sharpened by use and extended to the effective analysis of the behavior of complex informational systems prevalent in animals.

As a starting point, consider some of the ways in which one can elaborate on the standard notion of a transfer function, which forms the usual stock-in-trade of the servo-engineer and electronic designer when he is trying to understand the behavior of control-communication and computer systems. The transfer function is usually represented as a black box into which a signal is put and out of which a related signal emerges. Normally, it is very tightly hemmed in with formal restrictions of allowable behavior not unrealistic, in most cases, for amplifiers and physical transducers, but usually very unbiological. As a result, many a biologist has been led down the garden path when he has attempted to use this concept, for, unlike the engineer, he is unable to make his system conform to the limitations of the model on which he bases his reasoning.

In the conventional cases, a transfer-function black box will be classed as either passive or active. If passive, it will require that there be put in as a signal all of the power required to actuate the box. It merely gives out a proportion of this power modified in some internally determined manner. Alternatively, it may have a local power source, as in an amplifier, and thus actively give an enlarged or otherwise modified more-powerful or effective output. In either case, the transfer box is usually assumed to be linear, except for special cases. The output is, thus, in terms of magnitude at any one frequency, a kind of phase-shifted, differently scaled replica of the input, although the relationship may be different at different frequencies. For the simple cases then, the performance of the box is adequately represented by describing the ratio of the output to the input as a function of frequency when driven by a sine-wave input variation. Output is presumed proportional to input, but is acknowledged to vary in relative magnitude and also to advance or retard in relative phase as a function of frequency. Consequently, for many purposes, a complex plane phase-amplitude plot of this transfer function is considered to convey adequately the behavior of the box.

Usually, the transfer-function box is used as a building block to construct functional block diagrams, often including loops like those in Fig. 5(a) to confer a measure of stability absent in a system that is not back-compared. Occasionally, a long, compound loop system like that of Fig. 5(b) will be used, because it has still greater stability than the simpler loop systems for a given available gain, but it is also more likely to become violently unstable if certain limits are exceeded.

Note that the feedahead loop illustrated in Fig. 5(c), while not intrinsically as stable as the feedback loop, has an advantage in being capable of introducing exactly compensatory control corrections without having infinite gain, whereas the feedback loop can only approach this ideal asymptotically. Feedahead can be combined with feedback to give a very desirable compound operation and is now gaining widespread acceptance in engineering applications.

Figure 6 illustrates a more typical, biological-transfer loop system which characteristically includes feedahead as well as feedback loops and almost always displays control parameter dependency upon signal. It should be emphasized that the traditional transfer-function box clearly carries with it the notion of signal pathways firmly fixed spatially; that is to say, there is some kind of a discrete-circuit-pathway diagram characteristic of the system. There may be switches, but these merely direct activities between alternative discrete pathways. Each transfer box is a point-to-point function; whatever enters, comes in from a source at a specifiable place in the diagrammatic model, and what comes out, comes out at a specified place or places. Such function dia-

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

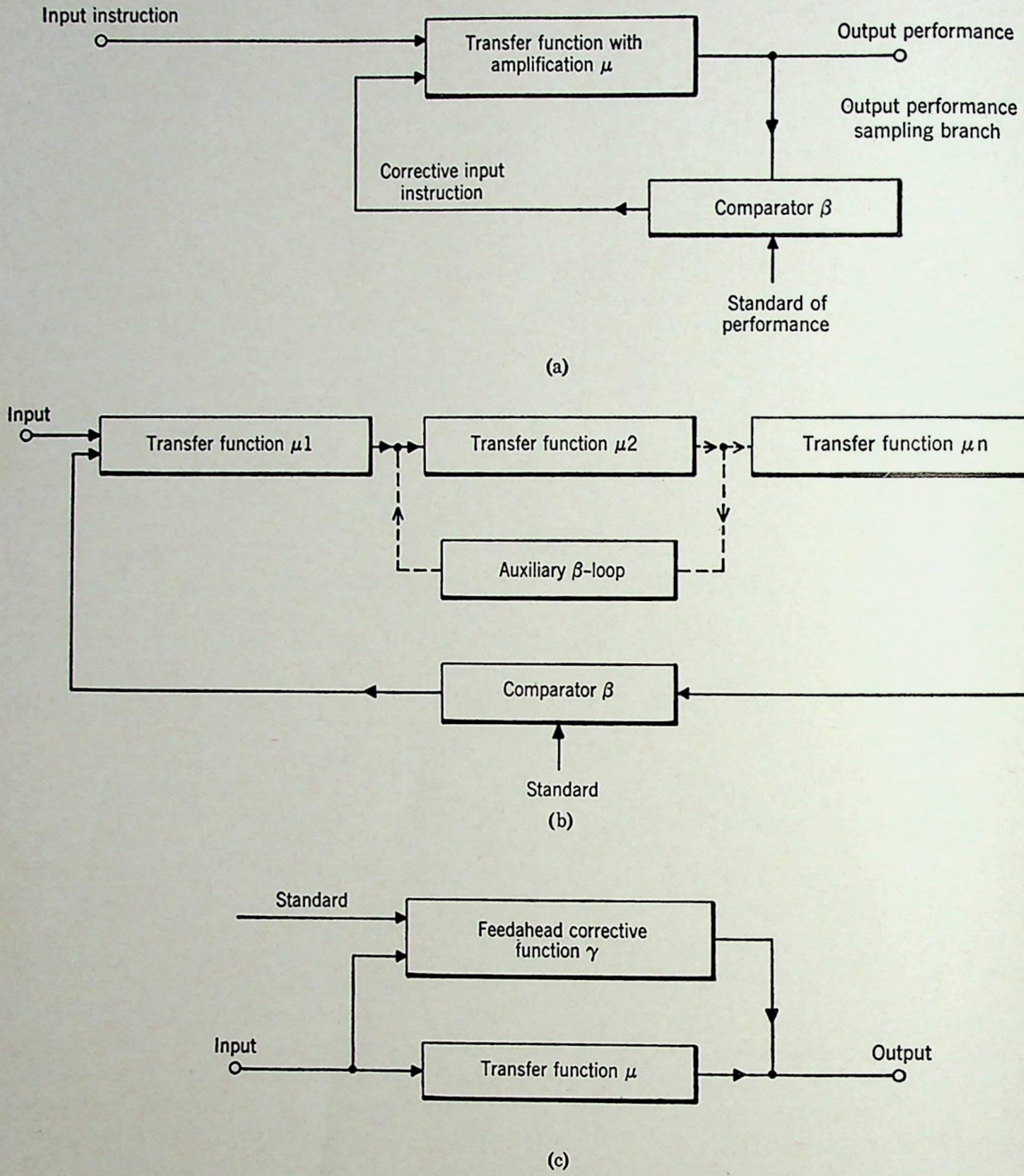


FIG. 5.

grams are hardly ever generalized to anything more complicated than a substantial interconnected block of such units, possibly with systems of switching arrayed in a logical, or occasionally in a statistically determined, pattern.

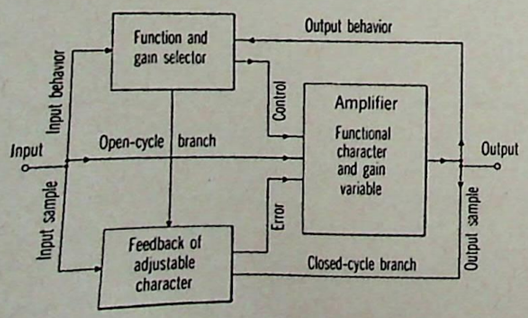


FIG. 6.

In all fairness, it should be pointed out that there is another and quite different set of packages, sometimes also called transfer functions, where the black box, instead of performing linear operations, now performs logical operations. These units, which figure prominently in digital-computer design, do something specific at an output when received signals in some predetermined logical set appear or fail to appear at its input. Typical signal sets might represent both, neither, or, *a* before *b*, time coincidence, equality, inequality, etc. These two quite different kinds of transfer-function black boxes can, in fact, be considered as epitomes, respectively, of

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

the functional organs of analogical-computer systems and of digital-computer systems.

Biology does not have *a priori* knowledge of a real difference between analogical and digital computers, and consequently is not hampered by this distinction. In fact, this difference is much more of the experimenter's making than it is an intrinsic property of the system being analyzed.

Digital systems are really those where enough probabilistic configurations of a system are lumped redundantly into a few significant categories to make the whole system approximate the all-or-nothing pattern characteristic of Aristotelian logical arguments. Analog systems are those where the available system configurations are dispersed over a significance range and, if numerous enough, resemble a value continuum. In the systems where energy redundancy is reduced to a minimum and nearly all of the available time band width of an information-handling system is utilized, the conventional thinking about digital and analog distinction becomes hazy and a new and more general type of analysis has to be substituted.

It is perhaps unfortunate that the term "analog system," which originally implied that one physical variable was being substituted for another for computational purposes, has come also to be used to describe any system utilizing a continuous range of variables in its operation. One is thus left with the anomaly that an analogical computer may be digital in makeup and that circuits performing quite direct control operations, and not particularly analogical physically to anything else, will be called analog-control systems.

It would be unfair to suggest that the modern engineer is not aware of the importance of nonlinear systems or to accuse him of thinking that digital and analogical systems are intrinsically different. He does, however, tend to build apparatus so as to conform essentially with one or the other of these because these analyses are familiar and easy, and he most likely has not been forced to cope with distributed transfer-function systems, particularly those that are highly re-entrant.

A little futher on, the behavior of two importantly different types of re-entrant systems is discussed, but first some places in biology where one can use fragments of available analyses are examined and embroidered on so as to make them more biological in their characteristics. In this way, they may be made applicable to the increasingly important cases where there are biological-transducer elements, particularly neural nets, which are simultaneously a continuum and a system of tight topological connectivities. This is a complicated way of saying that they are to an extent switchboard connected, but are simultaneously a system of diffuse field interactions. In this case, these field interactions mean simultaneously the chemical field, e.g., the hormonal or other chemical concentration-gradient field, and equally important, the electric field contributed by the surrounding active elements.

There will almost certainly be found in this complex an understanding of the efficient large-scale behavioral control of higher organisms, and mathematical techniques will have to be expanded to give facility and simplicity in handling systems which are hopelessly complex if treated as telephone exhanges gone berserk or soaked in electrically leaky saline.

One cannot offer more than an introduction to this kind of thinking at this time—first because it is complicated, and second because most of the answers are still unknown—but several extensions of ordinary field theory can be stated that are not presently being used in neural-net analysis.

There is one experimental and theoretical approach to this area that should be explored. This is the approach based on cardiac transfer-impedance theory, a technique used successfully in our laboratory in examining the superficially accessible action potentials of heart muscle, i.e., the spatial electrocardiogram. This theory, developed to handle biophysical heart data, will in all probability prove even more valuable in examining neural field interaction than it has in the purpose for which it was intended. Strangely enough, the key notion of reciprocity inherent in this theory was beautifully stated by Helmholtz, one of the early biophysicists, just about a century ago. It was not applied to biophysical problems until quite recently, and then it was introduced in another disguise.

The following demonstrates how the transfer-function notion can be spatialized. Consider Eq. (2), the definition of transfer impedance in its usual engineering form for a linear passive network:

$$\text{output voltage} \equiv \text{input current} \times \text{transfer impedance}. \tag{2}$$

Such a network may include any combination of capacitive, resistive, or inductive components that do not contribute external energy. For any combination of such linear components, one can determine the current applied between any two terminals in the system and the resulting potential difference developed between any other two terminals. The ratio of this voltage to the current will be dimensionally an impedance, and will have a constant value at any specific frequency if the system is linear.

Substitute for the passive circuit a biological tissue, nervous or otherwise, which is electrically at least approximately isotropic and linear although not necessarily homogeneous. These qualifications are included to avoid tensor notation with its complications when simple vector expressions are probably adequate for the present state of the art. In the low-frequency range which is of interest neurophysiologically, the brain approximates such a system of linear electrical components, except for microregions. Consequently, if any two points are picked as terminals through which to apply current, and any other two points are picked to

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

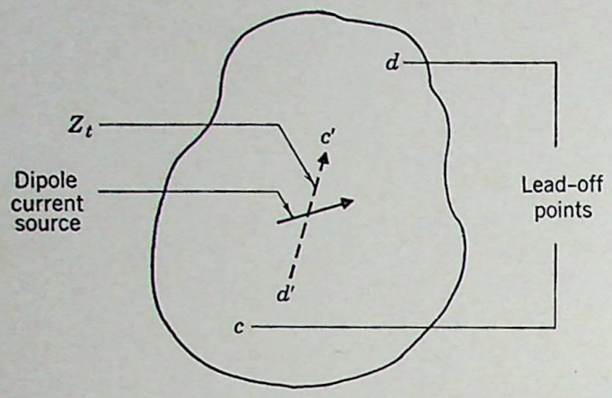


FIG. 7.

read off the potential difference, one finds that the voltage and current will be in a constant linear relationship which is formally described by their transfer-impedance ratio. If one now brings the two current points close together, they can be treated as a dipolar source-sink combination, and such a dipole is the kind of source that is encountered in biological tissue.

Because of the relatively short, length constants involved in muscle, nerve, and other biological current-generating systems, one can generally assume quite safely that source and sink are dipolar on a millimeter basis, and that macroscopic linearity prevails. In an isotropic medium, another generalization can be added. One can observe that, so long as the source dipole remains essentially in one region, it with its orientation becomes vectorial in the sense that the voltage transferred to a specified pair of terminals will obey a cosine relationship with respect to some specifiable preferred axis.

Thus, the dipole can be thought of as projecting itself, to a constant scale, on an effective axis not necessarily going through the recording points. In the diagram (Fig. 7), lead-off points c and d may be thought of as effectively represented by a vector on an axis $c'd'$ at the source. The magnitude of potential realized at the lead-off electrodes will be proportional to the magnitude of the current source times its distance of separation (i.e., its current dipole moment) multiplied by the cosine of the spatial angle between it and a reference line $c'd'$.

It is seen at once that this has become a vector dot product system and that the $c'd'$ reference has become a reference vector characteristic of a particular lead system. One, therefore, vectorializes Z in the alternating-current version of the Ohm's law relationship $E=IZ$. I becomes a dipole current moment (amperes times separation in centimeters). Z becomes a transfer impedance which is dimensionally ohms per centimeter in magnitude, but assumes the spatial attributes of a vector. Voltage, of course, remains a scalar quantity which is consistent with its representation as the scalar product of the transfer impedance with the current dipole moment.

$$V=\mathbf{I}_m\cdot\mathbf{Z}_t.$$

With Z vectorialized, one now has a function capable of being distributed spatially throughout a tissue system; that is to say, the transfer function has been successfully distributed in a neural field space. This function must necessarily be a vector-point function—that is, a quantity having specifiable magnitude and direction at every point throughout the space involved. By evaluating this function, one can characterize the interaction of a source anywhere in the field space with a receptive region anywhere else in the field, and can quantitatively predict their mutual electrical coupling.

It is at this point that Helmholtz's reciprocity theorem comes to bear. One must be rather careful in stating the theorem for a distributed medium, but it can be put in the following form. For any given system of lead-off electrodes, there is a spatial vector-point function which will relate the potential difference measured in this lead to the current dipole moment introduced anywhere in the medium in any orientation. The potential will be $V_t=\mathbf{I}_m\cdot\mathbf{Z}_t$ where V_t is the led-off voltage, $\mathbf{I}_m$ the dipole-source moment, and $\mathbf{Z}_t$ the transfer-impedance function. Upon applying an adaptation of the reciprocity relationship, it is found that the electric field or potential gradient $\mathbf{E}_t$ anywhere in the medium owing to introduction of current I into the former lead-off electrodes will be exactly $\mathbf{Z}_tI$, so that the spatial transfer-impedance function $\mathbf{Z}_t$ is identical in the two cases for any electrode or medium configuration.

What is the value of this $\mathbf{Z}_t$ function and its reciprocity property? In the first place, it permits one to predict, by applying currents to finite leads and measuring the resulting potential distribution with micro-electrodes, what potential will be set up as a field by microsources such as neurons. Model measurements and *in vivo* measurements of transfer impedances by the reciprocal method now become feasible and valuable. While these measurements have not yet been done to any great extent on brain tissues, it is possible to illustrate the kind of results to be expected by comparison with results obtained in the electrocardiographic field.

Figure 8 is an example of such an application where a set of orthogonal leads was developed experimentally by linear combination of partial derivatives of the transfer-impedance function. Knowing that transfer-impedance functions must obey the superposition theorem as well as the reciprocity rule, it was possible to develop leads sensitive, for practical purposes, *only* to sagittal, vertical, or horizontal sources of current irrespective of position within the heart tissue.

Inversely, it was possible to show theoretically and experimentally that fields from neurons in certain configurations can add up so as to be quite strongly concentrated at substantial distances from the sources, while in other cases they die out even more rapidly than the inverse cube law. Measurements of transfer-impedance fields via models give surprising insight into such complicated field patterns. The analysis indeed

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

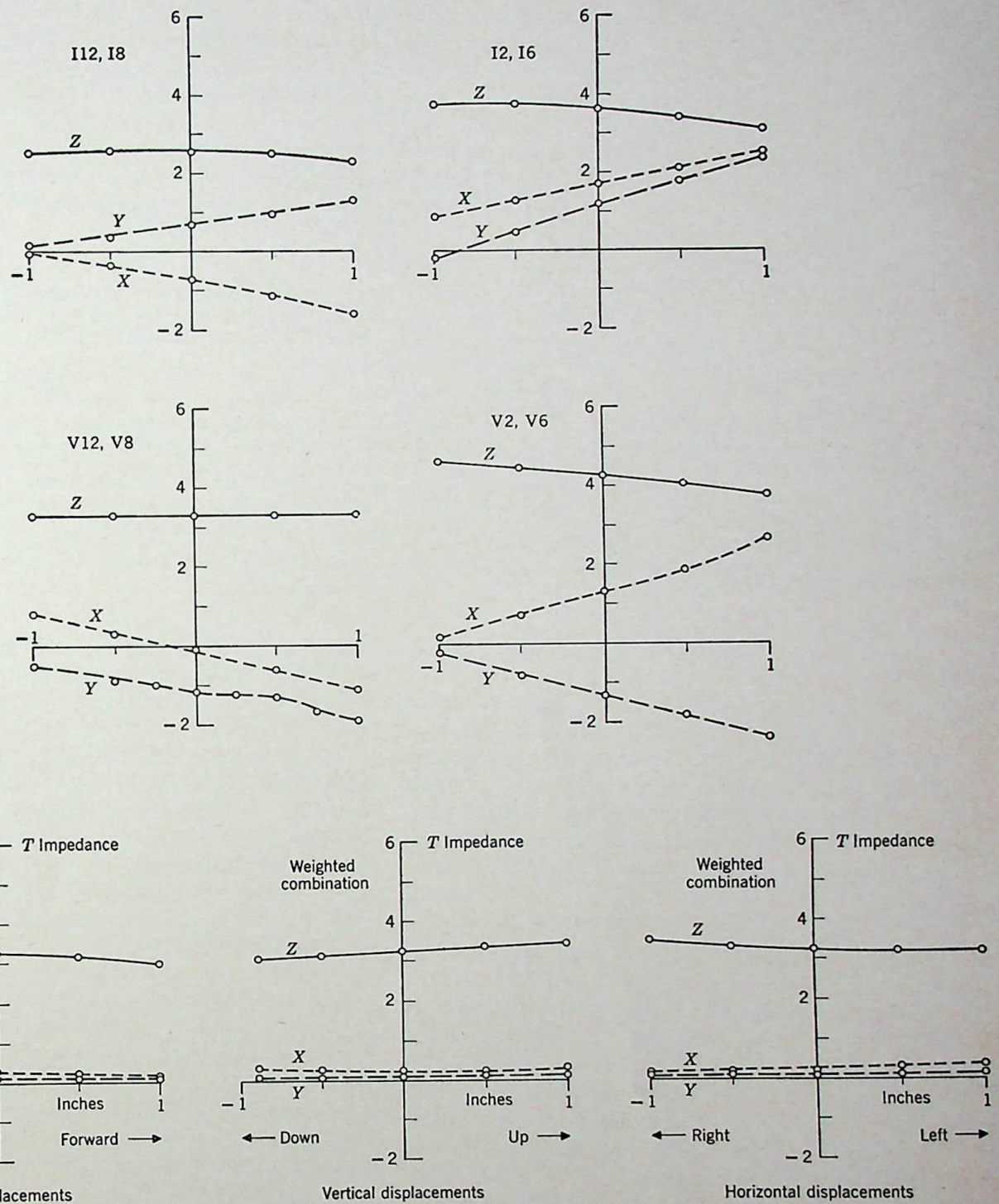


FIG. 8. Orthogonalization of transfer impedance fields. For any tissue system to which a specified lead system has been applied, there is a characteristic associated transfer-impedance field which takes the form of a continuous spatial vector-point function. By examining the principal partial derivatives of this vector function for several different lead systems applied to the same tissue, it is usually possible to devise weighted combinations of leads which yield nearly orthogonal functions in a finite source region so that potentials picked up from the combined leads report uniformly and orthogonally on sources in the corrected region, and stimuli applied through the leads stimulate the source region uniformly and in a constant known direction. Alternatively, highly distorting leads can be devised which will record and stimulate selectively. The patterns shown here as examples were developed to yield a uniform sagittal, or front to back, lead for the human chest to be used in electrocardiography. Ideally, the transfer impedance should be large and constant in Z and should have negligible X and Y contamination.

resembles a kind of degenerate directional antenna design procedure.

Spatial transfer-function analysis is very attractive in studying field stimulation of neurons, for it gives an intuitive insight into the remarkable time sensitivity of stimulation patterns. If one remembers that the conduction time at constant velocity is equivalent to distance, and that the inverse equivalence is also valid, then one has in a neural net a most interesting mechanism for the spatial-temporal localization of a pattern,

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

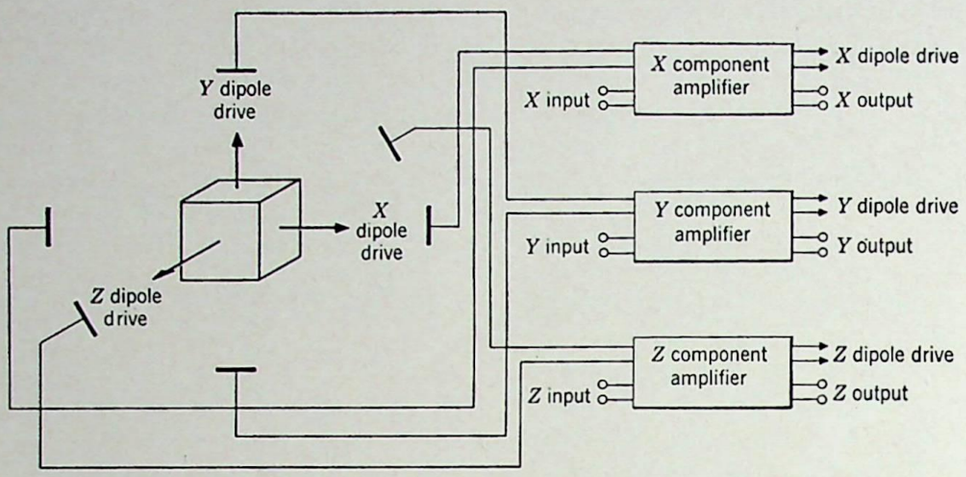


FIG. 9. System for automatic synthesis of equivalent concentrated dipole-current moment sources. A conductive model of the tissue system is fitted with a lead system equivalent to that on the experimental animal or person, and a vector dipole source is installed in the model at the chosen locus for which the equivalent vector source is to be calculated. A system of high gain amplifiers is looped around the model lead system to the dipole, and the amplified signal from the biological subject is inserted as a forcing function. In order to maintain all the amplifier inputs at near-zero potential simultaneously, the amplifier system must generate a dipole source which just cancels the complex of lead potentials and thus one which duplicates, in a negative sense, the desired source. The solution is, of course, unique only for a specified dipole source locus and electrode configuration as can be demonstrated by rearranging the electrodes on model and subject or by moving the dipole source about in the model and noting the continually altering resultant pattern of source dipole moment.

and hence an ideal medium for distributed probabilistic field effects.

Another closely related extension of transfer-function analysis was introduced by Burger in Holland a few years ago. In this approach, an image space is developed as a complementary space to the real field space. In the image space, vectorial summation of superimposed field ontributions becomes beautifully simple, even though orking out the shape of the image space is a more omplicated process.

There is a very interesting feedback method for synthesizing spatial excitation sources (Fig .9). If one builds a conductive three-dimensional scaled model of a tissue system and immerses in it an orthogonal, triple dipolar source at a place corresponding to a supposed bioelectric source, this artificial source will excite a transfer-impedance pattern similar to that of the original living system. If each component of the dipole is fed from a high-gain amplifier which is fed, in turn, from one of a set of three reasonably orthogonal lead sets corresponding to the dipole orientation, the system then becomes an automatic synthesizer of excitation sources in terms of the actual picked up potential. Explained in behavioristic terms, the feedback system tries constantly to null itself in all directions simultaneously and can do so only by creating a source equivalent to the real but inaccessible one. To what extent this general scheme can be extended beyond three loops to synthesize more complex sources, and to what extent noise can be used as a tracer to establish source localization, are not yet known, but this approach seems to be very promising.

Space does not permit amplification of many additional aspects of the biological-transducer function which are of great importance, but the transductive aspects of animal "orientation" behavior must not be totally omitted. Animal orientation includes especially the navigational processes whereby an organism "steers" itself toward a desirable goal literally as well as figuratively. It is well known that pigeons home, that fish almost unerringly find their native stream after months at sea, that birds migrate to very specific targets, and that bees communicate to their fellows the range and azimuth of food depots. Birds are even believed to utilize astronavigation, as suggested by recent experiments. All of these processes are biological-transducer problems at a higher level and ought to be investigated on this basis, as well as by the traditional types of experiment. If a moth knows which way to fly to find another moth a kilometer away, this is a transduction and coding problem, and one should seek not only the sensory transducers involved but also the transfer functions describing the reduction of massive sensory data to net decisive instructions for a flight plan.

Another set of biophysical phenomena classed as animal orientation is the family of rhythmic or oscillatory processes found in almost all biological organisms for which one has yet to discover a biophysical mechanism. It is well known that animals and plants exhibit strongly rhythmic properties with periods ranging from 1 or 2 msec in the case of some nerve processes to periods of a month, a year, or even several years in other cases. These rhythmic processes often have very nearly, but not exactly, a one-day period. Again, the period may coincide with the succession of high and low tides in a particular place, with the lunar month, or with the terrestrial year. While these periodicities are related to those of the normal environment, many of them persist for many cycles in experiments where the organism is carefully shielded from all known environmental clues to these periods.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

Some of the simple biorhythms like the cricket chirping rate have a reasonable chemical-process temperature coefficient but others have an almost exactly zero temperature coefficient. Some of these temperature-independent periodicities can be shifted in phase by a systemic shock such as dunking in cold water. Others can be "pulled," as a crystal oscillator can be pulled slightly off frequency by coupling to another resonator of similar frequency. None of the basic oscillatory mechanisms has been identified nor has an adequate hypothetical master oscillator been devised which will conform with experimentally determined properties of these natural oscillations.

Crabs change color on a tidal-periodic basis; some bacteria luminesce periodically on the basis of when it should be day or night; some flies emerge from pupae on an integral diurnal but not fixed illumination-period basis; some poisons change toxicity on a diurnal basis; and cilia beat rhythmically. No one seems to have found the transducer mechanisms by which these rhythms are loosely but adequately coupled to the environment. Surely, there should be a search for mechanisms of periodicity hinted at by the three major classes of physical oscillating systems. In any case, one should be embarrassed to have failed completely in turning up even approximate explanations for these phenomena that are of undoubted prime biological importance at every level from the intracellular to that of the entire organism.

Finally, there are two items referred to earlier—bivalent codes and hierachical re-entrant control systems. It is evident that biological control systems are, in general, organized in a series of semi-autonomous hierarchically arranged levels where each subordinate system left to itself will achieve a more primitive, less flexible, but still effectual, type of homeostatic control. Single isolated cells get on somehow in tissue culture, and a heart will beat in parts if it fails to receive coordinating excitations from its pacemaker. What are the mechanisms and the general control principles whereby several, often contradictory, courses of motor behavior are coordinated so as to remain unified and influenced by the several operator controls, yet not uselessly vacillatory or indecisive in action? Evidently, whole patterns of feedback and feedahead control are coordinatedly included or excluded from taking dominance in effective control. What does an animal do when exposed to unfamiliar, conflicting sensory information such as might originate from a minor malfunction of one of its transducer mechanisms? Generally, it chooses one set of the feasible interpretations of the available data and ignores incompatible parts of the others. Sometimes, as in the case of Von Holst's fishes (where the fish, used to having 1*g* gravity downward and light from above, are experimentally exposed to light from the side and to a centrifugally increased apparent *g*), they seek a weighted compromise. Very seldom do they fail completely, as do our rockets, from some single small malfunction.

In summary, one may conclude that there is at present no unified, formally manipulable system in the biophysical sciences for examining and describing systematically the functional characteristics of biological transducers and for studying their aggregate behavior when combined into functional ensembles. There is ample mathematical experience in the engineering fields to permit a start in this direction, but existing techniques fall far short of adequacy, and it is certain that new, uniquely biological theory and techniques for transductive-function manipulation and code specification will have to be developed.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

54

Initiation of Nerve Impulses in Receptor and Central Neurons

THEODORE HOLMES BULLOCK

Department of Zoology, University of California, Los Angeles 24, California

THE objective of this paper is twofold: to point out some of the problems which represent the present state of physiological analysis of some sense organs as detectors and transducers, and to point out a current view of the complex chain of events between transducing a stimulus and initiating a nerve impulse. It will be argued that this chain is similar in receptor and central neurons and that information about either one is relevant to the other.

Biophysicists long have been interested in sensory receptors, since they present in such a conspicuous form the conversion of physical events of the inanimate world into physiological events in the organism.[1] Some of the best-known achievements of biophysical science lie in this realm. Indeed, a vast literature exists on sensory capacities, analysis of the parameters of reception, and quantitative description of the subvarieties of unit receptors within modalities[2]. There continue to appear in the literature new kinds of receptors physiologically identified, and there remains a host of histologically known sense organs and diffuse receptors, differentiated visibly but not yet defined physiologically (e.g., pectines of scorpions, esthetes of chitons, osphradia of snails and clams, etc.).

This situation, virtually unique among organ systems, is not surprising if one thinks of behavior as the principal attribute in which the great wealth of animal types is differentiated, if one ascribes behavior to differentiation in the central nervous system and remembers the intimate and perhaps causal correlation between differentiation in the central nervous system and that in the sense organs. One has learned to expect startling discoveries in each new batch of journals—polarized-light detection, wind-speed indicators, gyroscopic-deflection sensing, infrared directional devices, olfactory separation of optical isomers, hydrostatic-pressure detection in supposedly gas-free organisms, ultrasound reception—not to speak of systems achieving a complex degree of analysis of signals such as the ultrasonic-pulse-reflection analysis in bats, the amplitude-modulation frequency analysis of the ears of crickets, the integration of proprioceptive information in the strike of a praying mantis, and of we do not know what information in many cases such as orientation of fish to turbulence and to other unseen fish. This list furthermore omits mention of eyes of various kinds and their fantastic achievements, of which Hartline writes (p. 515).

Since a systematic survey is patently impossible, I offer here only a small selection of some recent instances of progress in the *identification and characterization of receptors* with respect to their capacities and properties. A closer look is taken later at the *parameters of impulse initiation.*

ELECTRORECEPTORS

Recent work[3–6] has opened up a new sensory modality in the sensitivity to normally occurring electric fields, found in many species of fish of at least three unrelated families, the Gymnotidae, the Gymnarchidae, and the Mormyridae (Fig. 1). These fish have electric organs which emit pulses of low voltage—of the order of 1 v—in some species in brief, low-frequency bursts, and in other species continuously, hour after hour, with characteristic frequencies and pulse form and duration, diagnostic of the species (Fig. 2). The frequencies commonly lie between 60 and 400/sec, but in some cases exceed 1000/sec. The pulse duration—from the whole animal, representing activity of many electroplates—ranges from 10 msec to less than 0.2 msec. What is significant for us is that the behavioral evidence clearly shows in some species the use of these signals in orientation with respect to objects, apertures, and other emitting fish in the immediate environment. Clearly, the fish detects alterations in the pattern of the electric field in the water surrounding it (Fig. 3). Marked signs of agitation are elicited by a wire brought within range or by the discharges of other fish. Conditioned-reflex experiments show the ability to detect the presence of a stationary magnet outside the aquarium and to discriminate between conductors and nonconductors in the aquarium. The fish respond to the movement of a small electrostatic charge such as that produced by combing one's hair with a vulcanite comb. Sensitivity to imposed electric currents has been known, but what is of interest for us is the demonstration that a normally developed sensibility exists which is apparently used in nature to detect and to analyze in a complex way naturally occurring electric fields.

The sense organ is presumably the lateral line, heretofore regarded as a special form of mechanoreceptor. We will have to make room in our lists of modalities for a new category—electroreceptors. It is perhaps more curious that this modality is not more widespread. Muscle-action potentials can be recorded in the water at some distance from fish as well as from other animals, but so far the tests outlined in the foregoing, when applied to fish of other families than these three, have

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

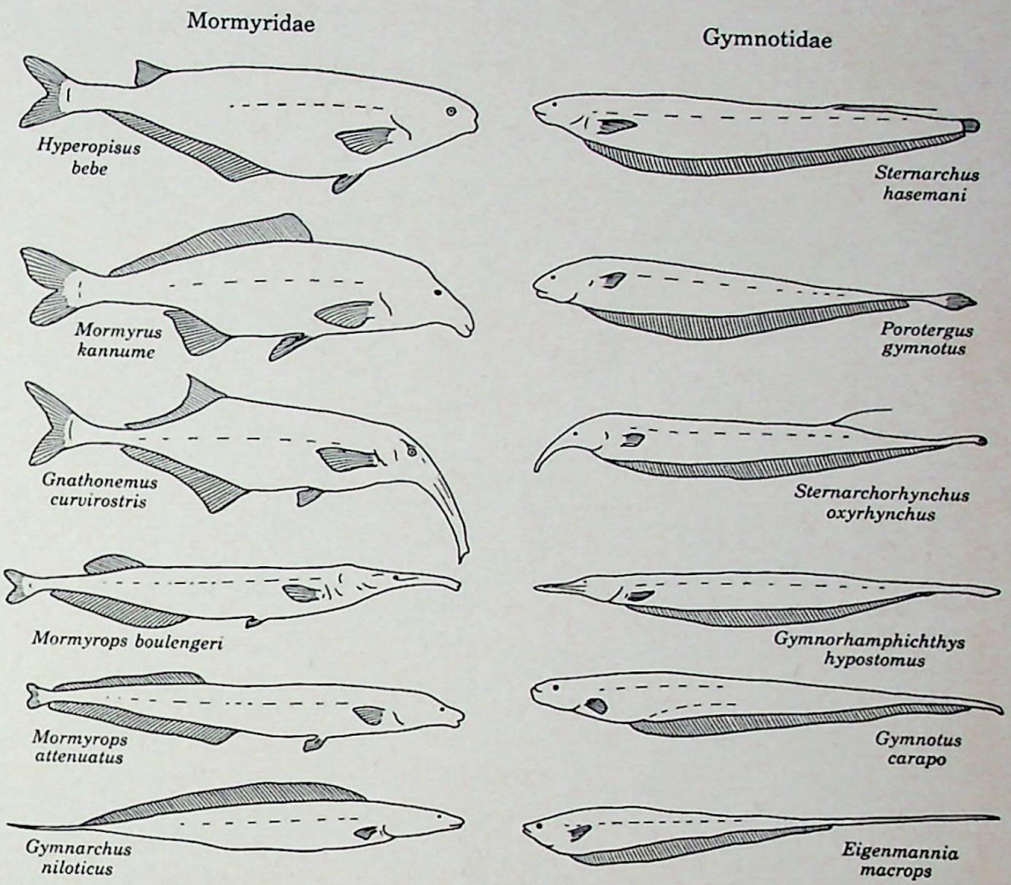


FIG. 1. Examples of continuously discharging, low-voltage electric fish. Representatives of the Mormyridae, Gymnotidae, and Gymnarchidae [from H. W. Lissman, J. Exptl. Biol. 35, 156 (1958)].

indicated no such sensitivity (with the possible exception of the *Siluridae*). One may conjecture that it is perhaps not so much the deficiency in peripheral sensibility as the lack of central apparatus for making use of the signals, which distinguishes the ordinary from the electric fish. All electric fish share, among other things, the enormous development of the valvulae cerebelli, the so-called mormyro-cerebellum.

The sensitivity of the receptor has not been measured directly but Lissman and Machin[3] have calculated the intensities of the electric fields which are effective in altering behavior, for example in conditioned response tests to electrostatic charges or magnets outside the aquarium. The field in the water around the fish must be changed by about 0.003 μv/mm at threshold, in *Gymnarchus niloticus*. This corresponds to a current through the fish of 2×10^{-5} μA/cm^2, some hundred thousand times smaller than a still subthreshold stimulus current density across the membrane of a squid axon (2 mv across 10^3 ohms/cm^2). The problem of the possible meaning of such high sensitivity to electrical events already has been raised by Terzuolo and Bullock.[7] Our evidence suggests that even in ordinary nerve cells the sensitivity, while not nearly as high as in these electroreceptors, is several orders of magnitude higher than the usually recognized threshold changes of 10 to 20 mv required to excite a silent nerve fiber. In the first place, this high sensitivity is manifested only as an alteration of frequency of an already discharging cell. In the next place, to be sensitive to minute changes in membrane potential, both the critical potential for spike initiation and the rate of rise of the prepotential must be extraordinarily stable—and must be localized in a limited part of the cell. The basic problem of high electrical field sensitivity—of special interest because it may be the one case where one does not need to search for a transducing mechanism as we do in photoreceptors, mechanoreceptors, etc.,—is the matter of *stability*. What are the requirements in terms of stability of the threshold and of the prepotential such that a given channel can provide a useful signal within a reasonable time

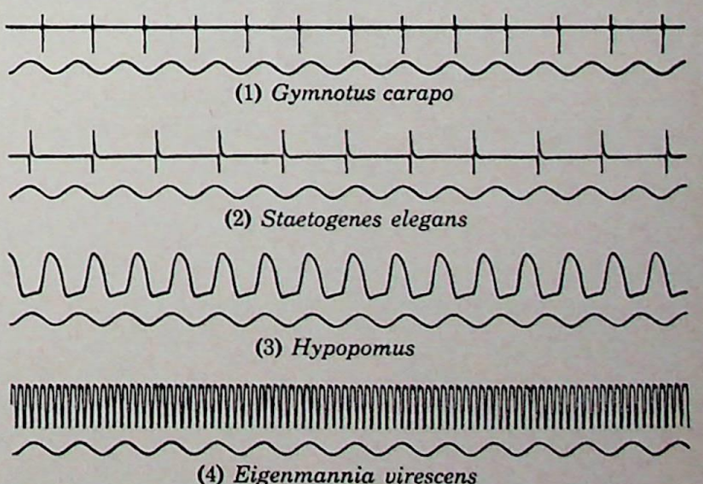


FIG. 2. Examples of types of electric pulses produced by four species of the Gymnotidae. Both pulse duration and pulse frequency are different and characteristic for each species. The fish are continuously discharging day and night. Time marker 50 cps [from H. W. Lissman, J. Exptl. Biol. 35, 156 (1958)].

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

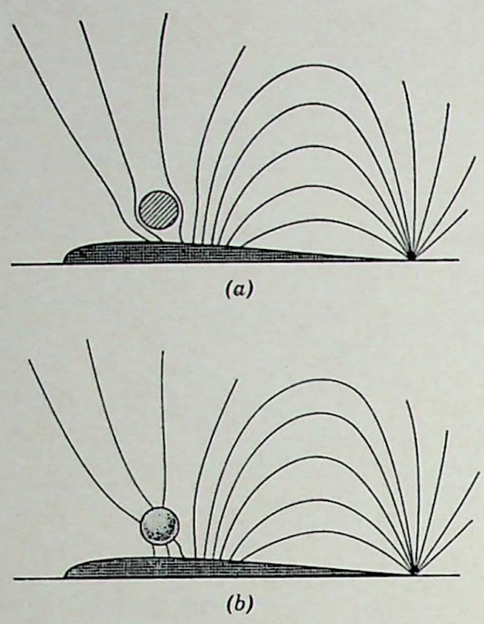


FIG. 3. The influence of objects upon the electric field around a fish. (a) An object of low conductivity, and (b) one of high conductivity. The fish detects the distortion of the field and reacts to both classes of objects [from H. W. Lissman and K. E. Machin, J. Exptl. Biol. 35, 451 (1958)].

(e.g., 1 sec)? At another level is the important problem of whether or not the high electric field sensitivity in the modulation of ongoing discharge plays a role in normal situations in the nervous system. If so, it permits effects not only of changes in electric fields normally present but also of changes in the chemical milieu which may act through the same mechanism, by changing the frequency of the spontaneous discharge for a given membrane potential at the critical locus.

MECHANORECEPTORS

Turning to the realm of mechanoreceptors, an attractive problem exists in the wide variety of arthropod hair-like exoskeletal projections (Fig. 4), each of which receives a single distal process of a primary sensory neuron and is specialized for signaling deflection of the hair in a certain direction.[8–10] In the lobster statocyst (an equilibrium sense organ where this is especially clear and for which recent analyses have been made), there are several specific types of hairs, differentiated for sensitivity to movement in one or another direction, or for sensitivity to a maintained position in one or another plane, or for sensitivity to vibration. Since the stimulus seems likely to be limited to a very restricted region—namely, the point of articulation of the hair near its base—one is faced with the general problem of transducing a mechanical event into a physiological one in perhaps a more simple and discrete form than in the more familiar mechanoreceptors such as stretch, touch, and auditory organs. Although the sensitivity must be high in terms of displacement, it cannot be estimated accurately, owing to the fact that, at the base of the long hair, where the relative movement that must act as a stimulus occurs, the mechanical disadvantage is

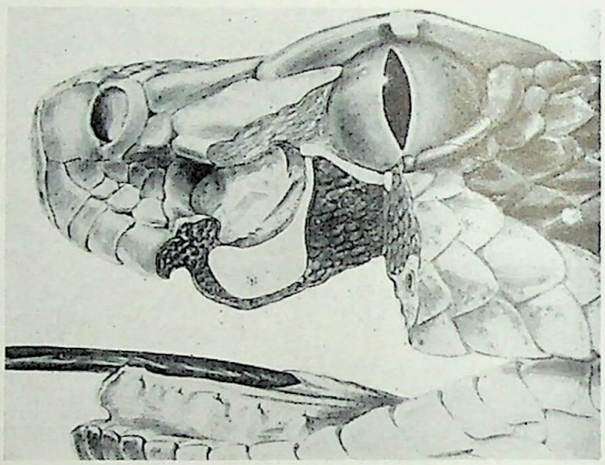


FIG. 5. The infrared sensitive facial-pit organ of rattlesnakes, *Crotalus*. Here, a wedge has been removed to show the deep pit and the sensitive surface—the thin membrane at its back (shown crinkled but normally smooth), with an air chamber both in front and behind it. This membrane, 10 to 15 μ thick, is heavily innervated and vascular [from T. H. Bullock and F. P. J. Diecke, J. Physiol. 134, 47 (1956)].

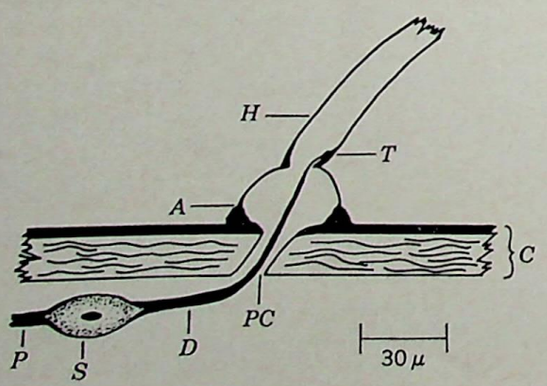


FIG. 4. Diagram of arthropod sensory exoskeletal hair of the type innervated by a single primary sensory neuron. The distal process of the sense cell ends near the region of articulation of the hair [from M. J. Cohen and S. Dijkgraaf in *The Physiology of Crustacea*, T. H. Waterman, editor (Academic Press, Inc., New York) (to be published)].

maximal and the small threshold movements of the tip of the hair must be reduced enormously in absolute magnitude. Fundamental questions remain, regarding the *ultrastructure and the local events* in the region of the mechanical deformation of the sensory nerve ending.

As an example of the specialization of which arthropod mechanoreceptors are capable, Burkhardt and Schneider[11] have found that units in the Johnston organ of the antenna of flies (*Calliphora*) are hardly less sensitive than the human ear to sound frequencies between 150 and 250 cps, the range of the wing-beat frequency. Even beyond this range the units follow the sound frequency faithfully, giving the animal sensory information corresponding to each wing beat with a delay of only about 1 msec. The speed of flight is controlled by this sense organ. For some other recent studies in this area, see references 12–17.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

FIG. 6. The sensory endings in the pit-membrane of a rattlesnake. This is a faithful drawing from a silver-stained whole mount of the 10 to 15 μ membrane. The nerve fibers end freely in palmate expansions with branching processes. 500 to 1500 such expanded endings occur per mm^2. Apparently all are of one functional type—so-called warm receptors. Width of picture, 150 μ [from T. H. Bullock and S. W. Fox, Quart. J. Microscop. Sci. 98, 219 (1957)].

TEMPERATURE RECEPTORS

Referring briefly to temperature receptors, there is some information regarding the extraordinarily developed long-infrared-radiation detectors found in rattlesnakes and other pit vipers in the thin membrane at the base of the facial pit (Figs. 5 and 6)[18,19]. This structure, richly provided with a special form of free nerve ending and specialized in other ways—such as, for the directional estimation of radiant sources—responds to very small doses, of the order of 10^{-11} small calories in $\frac{1}{10}$ sec on the area of the terminal ramification of one nerve fiber (2000 μ^2). The evidence indicates that this response is due to the change in temperature of the tissue, which both by indirect and direct methods is estimated to be of the order of 0.001°C, close to the value already found for the human.[20] Expressed as Q_{10}, the frequency of nerve impulses in a single fiber increases with temperature with a Q_{10} of about 10^{30}, a figure which offers considerable room for speculation about *high amplification with preservation of reasonable stability* (Fig. 7). Much more could be said about each of these cases with respect to physiological properties and characteristics of response. It is the purpose here only to call attention to the variety of opportunities and problems presented by a few recently studied receptors.

CHEMORECEPTORS

In the area of chemoreceptors, I choose only two recent reports. Hodgson and Roeder[21] have discovered labellar hairs in various flies in which two primary sensory neurons send distal processes to the chemoreceptive area at the tip of the hair. These two neurons have different modes of sensitivity: one responds only to sugars; the other to salts, acids, and alcohols. Not only is this preparation of interest because of the opportunity for the physiological *study of unit receptors of different chemical specificity*, but also, because of its additional sensitivity to mechanical stimuli and to tem-

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

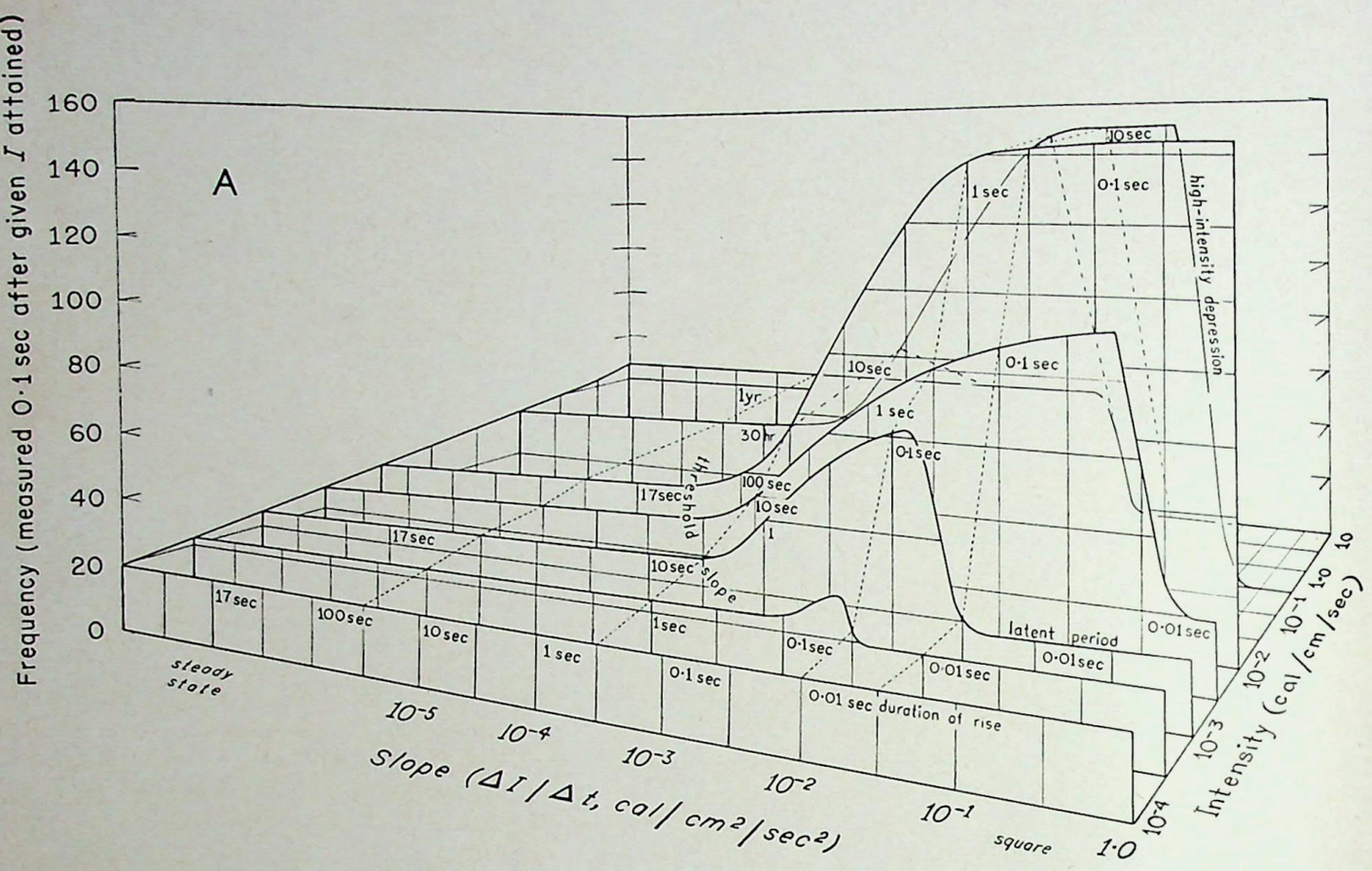


FIG. 7. Response of typical receptor unit of rattlesnake infrared sense organ, measured as nerve impulse frequency at different rates of increase of radiation and at different flux levels [from T. H. Bullock and F. P. J. Diecke, J. Physiol. 134, 47 (1956)].

perature changes as small as $\frac{1}{10}$°C, it presents in a clear form the general *problem of analyzing the world through receptors which are not unambiguously* detecting one aspect of the environment only. Furthermore, these labellar hairs are remarkable, as shown by Dethier,[22] in that complete behavioral rejection or acceptance responses of the intact animal occur to stimulation by a microdrop which can reach only one single sensory neuron. Schneider[23] has discovered an electrophysiological response in the antennae of silk moths (*Bombyx mori*) in the males only, which is highly specific apparently to the naturally occurring odorous material produced by the female. This odorous material attracts the males from great distances. In microelectrode records from the isolated antenna, picked up extracellularly from many units, Schneider finds a slow wave of several millivolts which is negative for some substances and positive for others. Spikes are superimposed preferentially on the negative phase of the electroantennogram. As in the preceding case, there is here an apparently peripheral filtering of the normally specific stimulus, which filtering is achieved by a *specific chemical sensitivity* of a receptor and again one is faced with high sensitivity. de Vries[24] has calculated that at threshold there are far fewer stimulating molecules impinging upon the sensory epithelium in man than there are olfactory receptor cells in the same area.

ABSOLUTE RECEPTION

Extending the remarks by Rosenblith (p. 485), it is highly desirable that further attention be given the problems raised by the sensory reception of absolute stimulus values as opposed to reception where comparison can be made by the receptors with a status in the recent past. Besides body temperature, blood pressure, and the like in man, there are many other indications of reception of values which may be called absolute, e.g., preferred temperatures in cold blooded animals, levels of light which day after day induce a given response, CO_2 concentration reactions, and pitch recognition. Even in many humans lacking phenomenal ability, the cross-modality subjective-intensity matching and many psychophysical phenomena illustrate scales not dependent upon relative stimulation as measured against a just-preceding level. One faces here the problem of stability and drift correction—not only the control loops, intra- or inter-cellular, but the standard of reference for detecting error.

SEQUENCE OF LABILE COUPLINGS LEADING TO IMPULSE INITIATION

Consider now what makes a neuron fire—namely, the specification of the distinguishable processes within the neuron leading to an explosive firing of a propagated all-or-none nerve impulse. The interesting conclusion

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

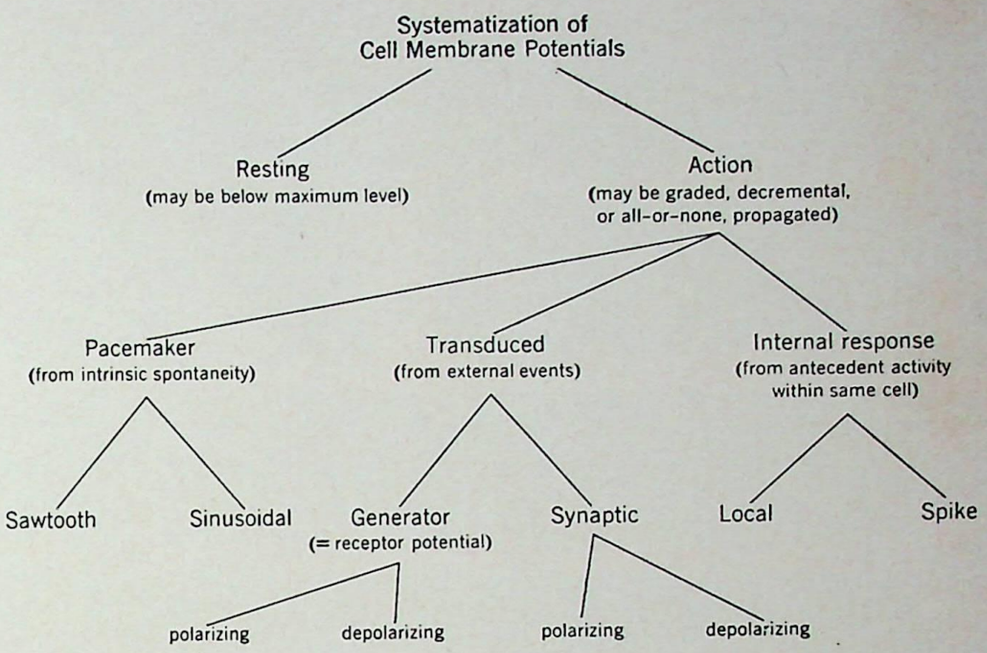


FIG. 8. The types of nerve cell membrane potentials.

permitted by current evidence is that *the determination of firing is not merely the build-up of an adequate stimulus* or its transduced and amplified resultant to a critical level, the threshold, but rather it is a sequence of labile couplings between graded (analog) events, each occurring in a limited fraction of the neuron. Normal firing is preceded by a series of steps, with alternative pathways. The several processes are separate, partly in space and partly in time, but causally are interconnected. Some of these processes are reflected in membrane-potential changes and some are not.

The *processes reflected in the membrane potential* include slow shifts of the average membrane potential, synaptic, generator, local, and pacemaker potentials. Figure 8 makes it clear that generator and synaptic potentials are parallel types of transduced potentials representing amplified response to external events whether from a presynaptic cell, from a non-nervous sense cell, or from the external environment. It also points out that both of these classes of response may occur either in polarizing or depolarizing directions. Depolarizing responses are commonly called excitatory because they are likely to increase the probability of firing. Polarizing, or as they are sometimes called, hyperpolarizing, responses generally decrease the probability of firing and are called inhibitory.

First among the processes which determine the initiation of impulses in receptor and central neurons, we should list what is actually no doubt a whole class of different and complex processes by which the level of the so-called *resting potential* is under the influence of the milieu and nonspecific agents (nonspecific to a given cell, though possibly specific to a cell type). Hormones, inorganic and organic constituents of the medium, deformation of the cell (even in neurons other than specific mechanoreceptors), and other factors are known to have some effects, although in most cases their importance in normal physiology cannot be assessed yet. In addition to these, we now can recognize as an important class the general shift in membrane-potential level accomplished through specific nervous pathways. For example, Otani and Bullock[25] found in certain cells of the 9-celled cardiac ganglion of lobsters that certain presynaptic fibers exert an influence, but without any discrete synaptic potentials. With repetitive stimulation they cause a slow, smooth shift of membrane potential (Fig. 9) as seen by an intracellular electrode in the soma. Terzuolo[26] has shown that certain neurons in the spinal cord of the cat similarly respond to stimulation of certain parts of the cerebellum by a shift of membrane potential level. This absence of discrete synaptic potentials even with a small number of incoming pathways possibly means that the main factor in determining the slow, smooth shift of potential is the distance of the synapses from the soma. If the distance is considerable, individual deflections would be smoothed out by the spread through leaky cables. But, if one believes that the membrane potential level at the basis of the axon or some such limited locus is crucial to the determination of cell firing, any such properties of the dendrites and soma—such as the electrotonic conduction of slow potential changes, the smoothing out of potential changes, or the discrimination against rapid deflections—would be of decisive importance in spreading the influence of the decrementally propagating activity of the much-branched dendritic processes to the region in which the spike originates. This may be one reason for having long dendrites. Among sensory neurons, there are many whose distal process divides to supply a considerable number of receptor regions of the periphery (Fig. 10, from the work of Meyer[27]) and in which the histology is suggestive that these loci are not alternative sources of full-fledged impulses but may, in some cases, make graded contributions to the probability of an impulse arising at some

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

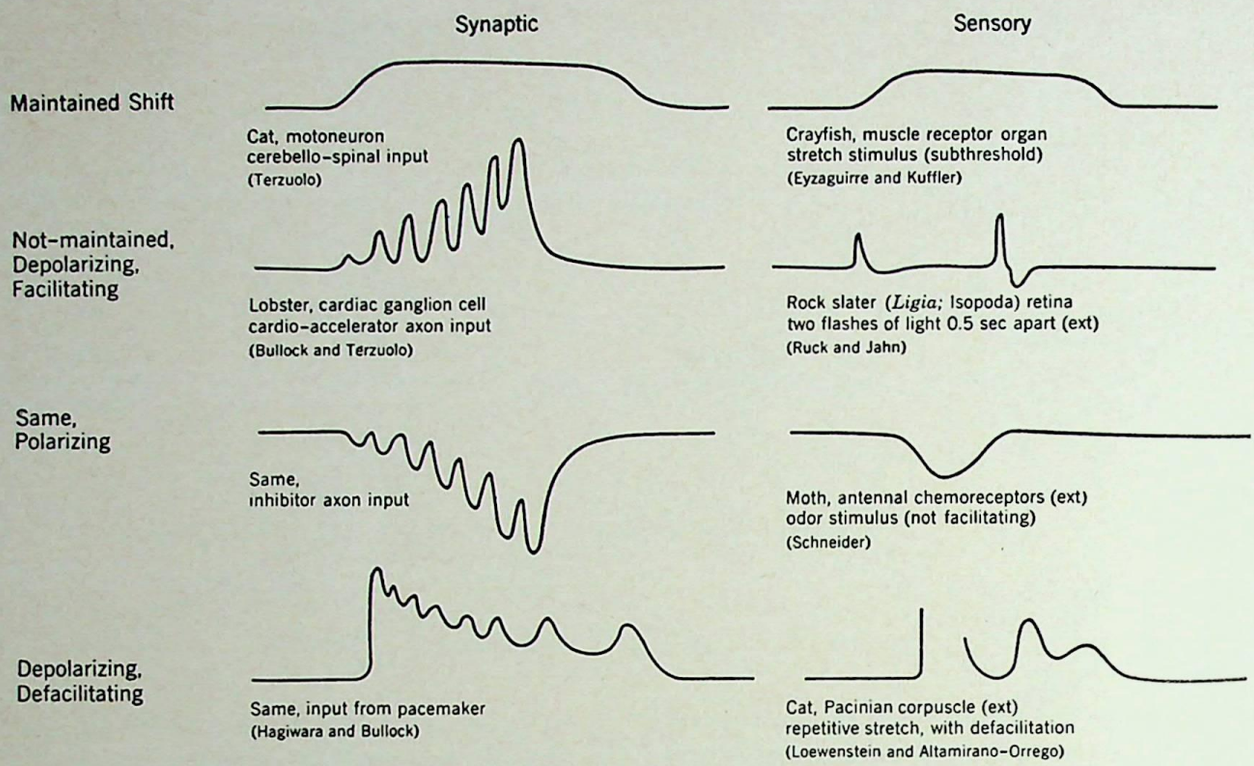


FIG. 9. Some of the types of subthreshold potentials, to show the correspondence between synaptic and sensory types. Hand-drawn approximations of original records, intracellular (except when marked ext = external); time scale is not uniform; upward deflection is depolarization or, in external recording, negativity of closer electrode.

point, such as a confluence of main branches of the distal process. So much for the processes that act upon the spike probability through the so-called resting potential.

More typically, or at least more familiarly, *the response to an external event is the transient deflection*, called a synaptic potential or a generator potential according to whether the external event is presynaptic or whether it is a sensory stimulus (Fig. 9). These responses arise in restricted portions of the neurons and have their significance in the effects they produce in other regions after nonregenerative spread—with decrement and phase delay. They may be polarizing or depolarizing, and this alternative may be decided by the nature of the external event or, in certain circumstances by the level of the membrane potential obtaining at the moment. Some normally polarizing, hence inhibiting, events are reversed readily to a depolarizing response which can be shown to increase the probability of firing and are thus at least in a measure excitatory. Such change in the response to a given input owing to different levels of membrane potential is not purely a laboratory phenomenon resulting from experimentally imposed polarization, but occurs normally, for example, in the aforementioned situation where Terzuolo found the membrane level of spinal motoneurons shifted by cerebello-spinal pathways, thereby altering the response to dorsal-root inputs. (In some accounts, emphasis has been laid on the fact that an inhibitory input drives the membrane potential from either side toward a certain equilibrium value and does not cause any synaptic potential if the membrane is already at that level. Even under the latter condition inhibition can result, because the lowered impedance tends to clamp the membrane. But this can inhibit only in the near vicinity. Upon spike-initiating regions or synaptic regions at some distance, the sign and magnitude of the potential change will determine the sign and magnitude of the effect.)

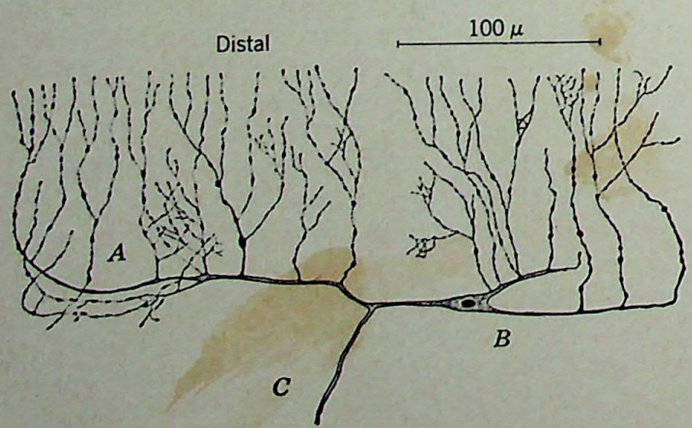


FIG. 10. Sensory nerve cell from the leg of a centipede, *Lithobius*; methylene blue. *A*—lattice of fine terminal fibers, *B*—cell body, *C*—axon [from G. F. Meyer, Zool. Jahrb. Anat. 74, 381 (1955)].

In at least some cases, there intervenes between the synaptic or generator potential and the critical level at the spike-initiating locus (which is a very restricted part of the neuron at some distance from the principal synaptic and generator loci) *an intermediate graded potential* which is usually called the *local potential* (Fig. 11, from the work of Bullock and Terzuolo[28]). This is presumed to arise in an electrically excitable membrane adjacent

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

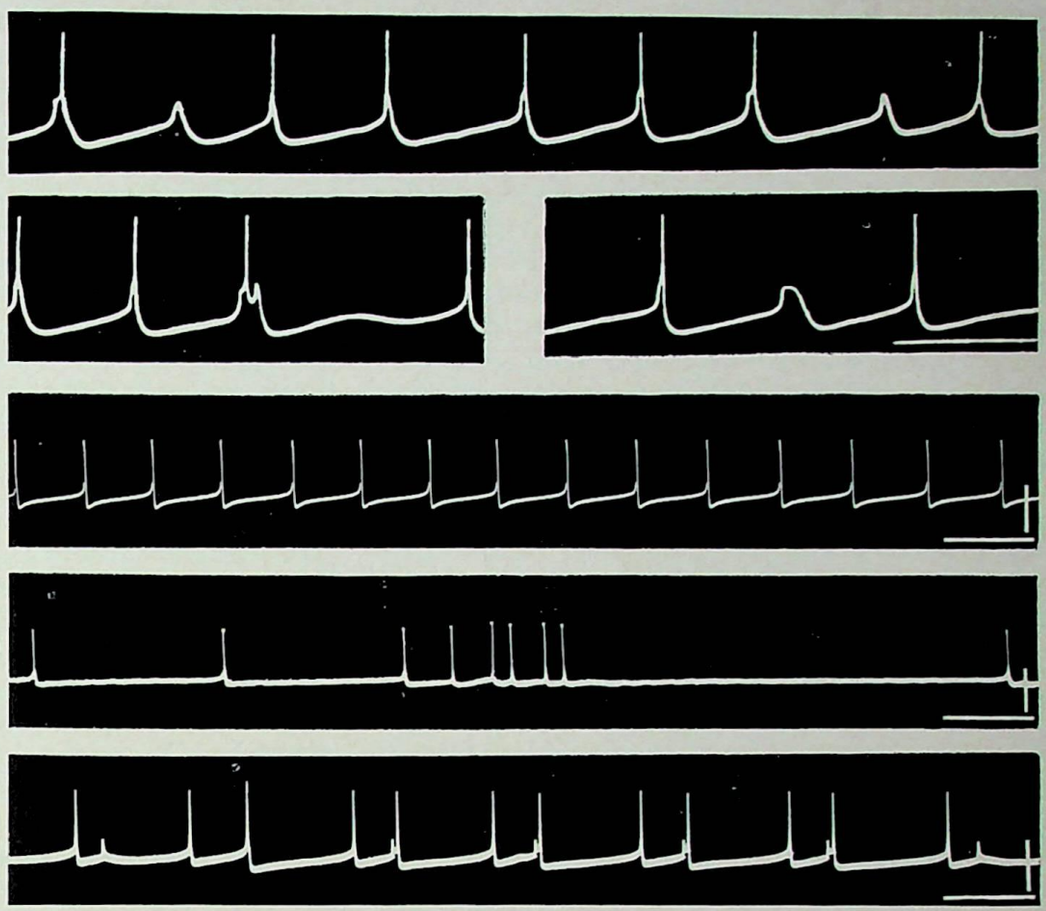


FIG. 11. Examples of true spontaneous activity as seen by an electrode inside a nerve cell. Cardiac ganglion cell of the lobster, *Panulirus* [(a) and (b)] or the crab, *Cancer* [(c) (d) (e)]. (a) and (b) show a large pacemaker potential, presumably arising nearby, and another prepotential, regarded as a local potential, before the spike. The local potential may fail to elicit a spike and then can alone cause repolarization. Note the failure of the local potential to arise following the third spike in (b) with instead an undulation leading to a new cycle. (c), (d), and (e) show different forms and permutations of pacemaker potential and repolarization. Scales: (a) and (b)—500 msec; (c), (d), and (e)—50 mv, 200 msec [from T. H. Bullock and C. A. Terzuolo, J. Physiol. 138, 341 (1957)].

to the synaptic or generator membrane, but distinct from it, and as a consequence of the local circuits from antecedent activity of the same cell. In this respect, the local potential is like the spike potential, which also arises by electrical excitation from the local circuits in neighboring regions of the same neuron. Indeed, local potentials can be made to have graded amplitudes all the way to the full amplitude of the spike potentials, and they differ only in being nonregenerative and, therefore, in spreading decrementally for distances of the order of a millimeter only. In some sensory neurons, we may conjecture that such local potentials occur in different branches of the distal process and sum in the stem process to determine the initiation of a spike. The experiments of Katz[29] on the summating subthreshold deflections in sensory terminals of muscle spindles may be so interpreted. Situations where distal sensory processes branch over a considerable area are quite common.

The nonregenerative property of some neuronal membranes is important in those cases where it results in the *inability of spike potentials, once initiated, to invade* those regions of the neuron. This preserves these regions from explosive depolarization and from subsequent strong repolarization, and thus enhances their integrative capacity. For example, the soma of the large cardiac ganglion cells in the lobster and perhaps the dendrites on some other types of neurons never experience an explosive all-or-none event. This characteristic is very likely true of most neuronal membranes, that is to say, of the vast forests of fine branching processes making up neuropiles. If this is so, one must think of all-or-none spike potentials as a peculiar development of a limited portion of the neuron whose function is the faithful propagation of signals over long distances rather than integration.

The familiar local and spike potentials are depolarizing in direction, but there are a few instances indicating that membranes are capable of equivalent *regenerative potentials in the opposite direction.* For example, during the plateau of an essentially completely-depolarized potential in the Purkinje tissue of the mammalian heart, during a heart beat, a threshold stimulus in the polarizing direction sets up a regenerative repolarization which grows explosively and restores a resting potential [Weidmann[30]]. Hisada[31] has recorded action potentials of the same polarity—that is, increasing polarization across the cell membrane of the protozoan *Noctiluca.*

Finally, we have *spontaneous activity* manifested in *pacemaker potentials* in some neurons including many sensory neurons. A background of continuous spike discharge in the absence of known stimulation or under steady-state stimulation may be called spontaneity. Of course, such activity depends upon permissive conditions of the metabolism and of the milieu of the cell. If these conditions are normal ones, however, steady-state firing is spontaneous in the sense that the origin of the intermittent activity lies within the neuron, rather than

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

in the environmental mechanisms. Such activity has been explored in several cases (Fig. 11), and the significant finding from the present point of view is the *localization of the pacemaker process to a limited portion of the neuron* different from that in which spikes arise, and perhaps also different from that in which synaptic or generator activity occurs primarily. Of course, these regions must be within shouting distance of each other, for their normal significance lies in the summing of their effects to produce a threshold change in the membrane potential at the spike-initiating region. In lobster cardiac ganglion cells, evidence of more than one pacemaker locus is seen in the same neuron. This is possible because these loci apparently occur in different major processes of the cell, each of which has its own spike-initiating region; two different rhythms of spikes arise and can be seen in the same soma without interfering one with another since they cannot invade the soma. What this says is that the absolute dimensions of neurons are not accidental but are fixed by engineering requirements specified by the electrotonic parameters of the membrane.

There are very likely at least *two distinct types of spontaneous potential* change. The rarer is approximately sinusoidal. Spikes tend to occur during the phase of maximum depolarization. Several spikes may occur during this phase, but they are not necessary for and do not accelerate the repolarizing phase. In contrast, the more common type is similar to a relaxation oscillation, in that the pacemaker potential itself is a steadily depolarizing potential which must be interrupted by some new process with a threshold whose net result is to repolarize, thereby starting a new cycle. This threshold process is typically the spike potential, but it may be preceded by and indeed may be substituted by a local potential which is capable of repolarizing the membrane during its recovery phase, thus setting the stage for another pacemaker potential to begin.

There are other *permutations and complexities in the processes of spontaneous impulse initiation* as is suggested by the high rhythmicity of some receptors and the low rhythmicity of others. In fact, it may be characteristic that, in one and the same unit, rhythmicity is relatively high at high frequencies of discharge and low at low frequencies. Tokizane and Eldred[32] have found two distinct populations of stretch receptor fibers in the dorsal roots of cats. One population is consistently more rhythmic at any given average frequency than the other; that is, the standard deviation of intervals is smaller for a given average interval. They believe that the more-regular units come from flower-spray endings in the muscle spindles and that the less-regular units belong to annulo-spiral endings.

The various possibilities for understanding the *origin of nonrhythmicity* or randomness-within-limits of successive intervals cannot be discussed here. Nor can one discuss the interesting question of the central problem created by nonrhythmicity, the *distinction of signals from noise*—that is, of weak true stimuli from random changes in the frequency of firing. These problems have been discussed elsewhere (see references 33–35).

These separate processes reflected in membrane potential shifts, in synaptic, generator, local, pacemaker, and spike potentials are *sequentially coupled* in complex ways: partly because there are several alternative sequences; partly because much depends on the particular microanatomy of a given neuron, the spatial relations of the separate loci, and the possibilities of spread of the respective potentials; and partly because of the labile character of the coupling constants. The constants or transfer functions well may be nonlinear.

However, these do not exhaust the processes of which there is evidence that determine the initiation of impulses within the cell. Turning to *processes not manifested by membrane potential* changes—at least by changes detectable with the same methods used for measuring the just-discussed events—we are faced mainly with indications of excitability changes. These are brought to attention only by responses to subsequent stimuli. One is called *facilitation* (Fig. 9). The cell can be initially at a certain membrane potential and respond to a given arriving impulse (in this case, a single presynaptic fiber) with a small synaptic potential. After repetition, the response to the same presynaptic impulse is much greater even when the membrane potential is allowed to return to the initial value between stimulations. Another type, which may be seen in the same neuron may be called *defacilitating*; the response is less to successive presynaptic impulses the closer together they come. Here also, at the same membrane potential level, the cell or presynaptic terminals are altered in response or excitability according to the recent history. This is not a laboratory artifact but is an essential part of the normal mechanism of burst formation in the 9-celled cardiac ganglion of the lobster. Here, the large follower neurons, driven by a burst of arriving presynaptic impulses from the pacemakers, characteristically give the form of synaptic potential response shown in Fig. 9. A large, initial synaptic potential is followed by a series of small ones during the high-frequency portion of the burst and then by growing amplitude responses as the frequency of the arriving presynaptic impulses declines.

Here we have to deal with the *time course of excitability of graded, subthreshold events*. There must be a separate curve for each of the subthreshold events previously discussed and for each locus and type of synapse. The last two instances, of facilitation and defacilitation, recorded in the same neuron were both excitatory synaptic potentials, from different presynaptic pathways (one from the pacemaker in the cardiac ganglion, and the other from the cardioacceleratory nerve from the central nervous system). The same neuron also has facilitating inhibitory synapses. Each of these synaptic

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

responses must have its own dependence of response upon input and of this relation (response per input) upon time after preceding activity. The local potentials also may have a separate time dependence of excitability and so may the spike-initiating locus where the excitability is measured by a sharp threshold. We note that the classical spike threshold, so easily measured any place in the axon, is normally a significant parameter at a very limited region only in the whole neuron. Beyond this region, the safety factor is normally sufficiently high that the spike threshold is without great significance.

Another degree of freedom not predicted by the membrane potential is the presence or absence of after-effects —either persistent response after the input ceases or of overshooting rebound (post-excitatory inhibition or post-inhibitory excitation). Cases of each kind are known for receptors and for central neurons.

Still another line of evidence for processes determining spike initiation, which cannot be seen in the usual membrane potential measurements, is that of significant *alteration of the frequency of already-active neurons by weak electric fields* in the surrounding medium, already mentioned. In a preliminary analysis, Terzuolo and Bullock[7] estimated the intensity of the voltage gradient in the saline surrounding a neuron (stretch receptor of the crayfish) when an imposed polarization was of a magnitude just sufficient to cause a noticeable change in the maintained frequency of firing. The results, as already mentioned, were that very weak fields (of the order of 1 $\mu v/\mu$ in the external medium around the cell) were sufficient. There will be a potential change in a given direction (for example, an excitatory direction) in certain parts only of the neuron, and since this will be graded in magnitude geographically, there must be a very limited region only wherein the imposed field actually exerts its effect. It can be concluded, therefore, that, even in the absence of average change in membrane potential, very localized regions may be critical in determining the firing frequency and may have an extremely high sensitivity to small voltage gradients. Thus, old ideas are confirmed in a certain sense that, among other things, the neuron is to a significant degree under the influence of differences of potential between one part and another of the surface of the neuron.

The general picture then is quite different from that of the axon. Grundfest[36] has given arguments for believing that adjacent patches of neuronal membrane are, respectively, electrically excitable and electrically inexcitable in a certain meaning of these terms. It may be emphasized that the evidence cited here is also strongly indicative of a *neuronal membrane consisting of a patchwork* of different kinds of subthreshold-response capacities. The responsiveness or the excitability—these two are different, but ordinarily are not easy to distinguish experimentally—vary separately one from another; some are manifest by electrical events, others are not.

One more degree of complexity must be added: the separate processes do not simply sum algebraically to determine the level of probability of impulse initiation, but at least some of them apparently interact: the presence of an increased magnitude of one subthreshold process may alter the magnitude of one of the others. For example, in the presence of a large generator potential, excitatory synaptic potentials may not simply add but will themselves be altered in size. The same may be true for inhibitory and pacemaker potentials. These interactions have been little studied, but there is no doubt that they will be found, in some cases, to contribute importantly to the integration of input to determine output (see, for example, the evidence of multiplicative neural events adduced by Hassenstein and Reichardt[37]).

SUMMARY

A selection of sense organs upon which recent progress has been made is reviewed in order to point out problems and opportunities of biophysical interest. The electroreceptors in electric fish, mechanoreceptor hairs in arthropods, infrared receptors in pit vipers, and specific chemoreceptors in flies and moths are the examples presented.

The general problem of absolute reception as opposed to reception of change is emphasized in terms of stability, drift correction, and reference standard.

The long series of subthreshold events which may intervene between transducing a stimulus and initiating a spike is reviewed. These include events which are reflected in the membrane potential and others which are not. Specific shifts of membrane potential, pacemaker, generator, synaptic, local and spike potentials, and some subvarieties are sequentially coupled in labile and perhaps nonlinear ways. They occur, in general, in restricted regions of the neuron which interact in complex ways not only because of the complex coupling constants between successive steps but also because of the profound influence of the anatomical distribution of the differently responding types of membranes.

Events not reflected in the membrane potential are the excitability cycles and the presence or absence of aftereffects. In addition to the spike threshold, each graded process preceding the spike has a curve of response against input and of this relation against time after preceding activity.

It is probably important that these several processes not only are sequentially related but that they interact —the amplitude of one may alter the responsiveness of another. Spike initiation is, therefore, potentially a highly derived and integrative result.

ACKNOWLEDGMENTS

Permission to reproduce figures from the *Journal of Physiology*, *The Journal of Experimental Biology*, and *The Quarterly Journal of Microscopical Science* is gratefully acknowledged.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

BIBLIOGRAPHY

[1] R. Granit, *Receptors and Sensory Perception* (Yale University Press, New Haven, 1955).

[2] T. H. Bullock, Federation Proc. **12**, 666 (1953).

[3] H. W. Lissman and K. E. Machin, J. Exptl. Biol. **35**, 451 (1958).

[4] H. W. Lissman, J. Exptl. Biol. **35**, 156 (1958).

[5] H. Grundfest, Progr. in Biophys. and Biophys. Chem. **7**, 1 (1957).

[6] C. W. Coates, M. Altamirano, and H. Grundfest, Science **120**, 845 (1954).

[7] C. A. Terzuolo and T. H. Bullock, Proc. Natl. Acad. Sci. U. S. **42**, 687 (1956).

[8] M. J. Cohen, J. Physiol. **130**, 9 (1955).

[9] M. J. Cohen and S. Dijkgraaf in *The Physiology of Crustacea*, T. H. Waterman, editor (Academic Press, Inc., New York) (to be published).

[10] C. A. G. Wiersma and E. Boettiger (to be published).

[11] D. Burkhardt and G. Schneider, Z. Naturforsch. **12B**, 139 (1957).

[12] C. Eyzaguirre and S. W. Kuffler, J. Gen. Physiol. **39**, 87 (1955).

[13] L. H. Finlayson and O. Lowenstein, Proc. Roy. Soc. (London) **B148**, 433 (1958).

[14] E. Florey, Z. Naturforsch. **10B**, 591 (1955).

[15] E. Florey, Z. Naturforsch. **11B**, 504 (1956).

[16] S. W. Kuffler in Submicroscopic Organization and Function of Nerve Cells, H. Fernández-Morán and R. Brown, editors, Exptl. Cell Research Suppl. 5, 493 (1958).

[17] W. Loewenstein and R. Altamirano-Orrego, J. Gen. Physiol. **41**, 805 (1958).

[18] T. H. Bullock and F. P. J. Diecke, J. Physiol. **134**, 47 (1956).

[19] T. H. Bullock and S. W. Fox, Quart. J. Microscop. Sci. **98**, 219 (1957).

[20] J. D. Hardy and J. W. Oppel, J. Clin. Invest. **16**, 533 (1937).

[21] E. S. Hodgson and K. D. Roeder, J. Cellular Comp. Physiol. **48**, 51 (1956).

[22] V. Dethier, Quart. Rev. Biol. **30**, 348 (1955).

[23] G. Schneider, Z. vergleich. Physiol. **40**, 8 (1957).

[24] Hl. de Vries, Progr. in Biophys. and Biophys. Chem. **6**, 207 (1956).

[25] T. Otani and T. H. Bullock, Physiol. Zool. (to be published).

[26] C. A. Terzuolo, (personal communication, 1958).

[27] G. F. Meyer, Zool. Jahrb. Anat. **74**, 381 (1955).

[28] T. H. Bullock and C. A. Terzuolo, J. Physiol. **138**, 341 (1957).

[29] B. Katz, J. Physiol. **111**, 261 (1950).

[30] S. Weidmann, Ann. N. Y. Acad. Sci. **65**, 663 (1957).

[31] M. Hisada, J. Cellular Comp. Physiol. **50**, 57 (1957).

[32] A. Tokizane and E. Eldred, (personal communication, 1958).

[33] T. H. Bullock in *Recent Advances in Invertebrate Physiology*, B. T. Scheer, editor (University of Oregon Press, Eugene, 1957), p. 1.

[34] R. Fitz-Hugh, J. Gen. Physiol. **40**, 925 (1957).

[35] R. Fitz-Hugh, J. Gen. Physiol. **41**, 675 (1958).

[36] H. Grundfest, Physiol. Rev. **37**, 337 (1957).

[37] B. Hassenstein and W. Reichardt, Z. Naturforsch. **11B**, 513 (1956).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

REVIEWS OF MODERN PHYSICS VOLUME 31, NUMBER 2 APRIL, 1959

55

Receptor Mechanisms and the Integration of Sensory Information in the Eye

H. K. HARTLINE

The Rockefeller Institute, New York 21, New York

THE processing of sensory information begins in the sense organs themselves. It is in them that the first steps take place in the transformation of external influences into the patterns of nervous action that regulate the activity of an animal in its complex environment. The fundamental nature of the receptors and the design of the accessory structures in which the receptors are deployed determine how external information flows into the organism. Thus, sense organ and receptor mechanisms determine the character of the neural activity that is passed on to higher neural centers. In addition, the first steps in neural integration take place within the sense organs, for in many of them the receptors interact with one another. As a result of both of these actions, patterns of sensory nerve fiber activity transmitted to the higher centers are more than mere replicas of the temporal and spatial patterns of external stimuli. Certain significant features of the stimulus patterns are accentuated at the expense of less important fidelity of representation. This can be clearly illustrated in the analysis of the first steps of the visual process, with which this paper deals.

One of the great contributions of biophysics in the last century was the precise description of the human eye as an optical instrument. The high degree of perfection of our eyes enables us to exploit many of the peculiar advantages of luminous energy as a source of information; their shortcomings set limits to our visual performance. The vertebrate scheme of optical imagery by a lens system is not the only one that is used by animals; compound eyes also have been evolved—made up of small optical units, each having a narrow entrance angle and each pointed in a different direction so that all cover the entire field of view. They too have both advantages and disadvantages, one of the advantages being that short wavelengths can penetrate to their receptors. In either case, retinal receptors arranged in a mosaic receive light in varying amounts from the various parts of the animal's surroundings. The mechanism of the visual receptor units that compose the retinal mosaic determines many of the properties of vision.

The photoreceptor offers certain advantages over many other receptors in the study of sensory mechanisms, for in it the very first step in the transducer mechanism for translating the stimulus into nervous action is beginning to be well understood. This is a consequence of the general principle that electromagnetic radiation, to produce a permanent effect on a material system, must yield some of its energy to the system. Consequently, the action spectrum of the visual apparatus is simply the manifestation of the absorption spectrum, or a portion of it, of the primary photosensitive material in the visual receptors. It is the absorption spectrum of the primary visual pigment that sets the rather indistinct limits to the extent of the visible region within the electromagnetic spectrum and that determines quantitatively the relative effectiveness of different wavelengths of visible light. There is now very good agreement between the measurements of the absorption spectrum of photolabile pigments extracted from the retina and the "action spectrum" of vision for several animal forms, especially for man.[1,2]

The fact that one can identify the photosensitive material of the visual receptor makes it possible to say whereabouts in the receptor cell the first act of the visual process takes place. In the vertebrate eye, the visual pigment "rhodopsin" is known to be concentrated entirely in the outer segments of the retinal rods. This identifies the outer segments as the locus of the initial step in the visual receptor process. Rhodopsin can be extracted from suspensions of the outer segments of retinal rods, and its absorption spectrum, after appropriate correction, agrees well with the distribution of spectral sensitivity of rod vision. It is clearly the primary photosensitive substance of the rods. A number of visual pigments related to rhodopsin are now known. One of them, iodopsin, is the corresponding photosensitive substance of the retinal cones.[3] The biochemistry of the visual pigments constitutes an extensive and elegant chapter of modern biochemistry that cannot be discussed in detail in this paper (cf. Wald[4]).

Rhodopsins are known to be conjugated proteins, the prosthetic group being a carotenoid called retinine. Retinine is an aldehyde, the corresponding alcohol being vitamin A, and is known in a number of isomeric forms. The first act of light apparently is to produce an isomerization of the carotenoid group while it is still attached to the protein.[5] Retinine is then split off the protein molecule by subsequent reactions that are independent of light, and may be converted reversibly into vitamin A.

After photolysis, visual pigments can regenerate. Otherwise, one would have one look at the world and then be forever blind. The kinetics of photolysis and regeneration of visual pigments has been studied extensively, both *in vitro* and recently in the living retinas of experimental animals and human subjects.[6] Many

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

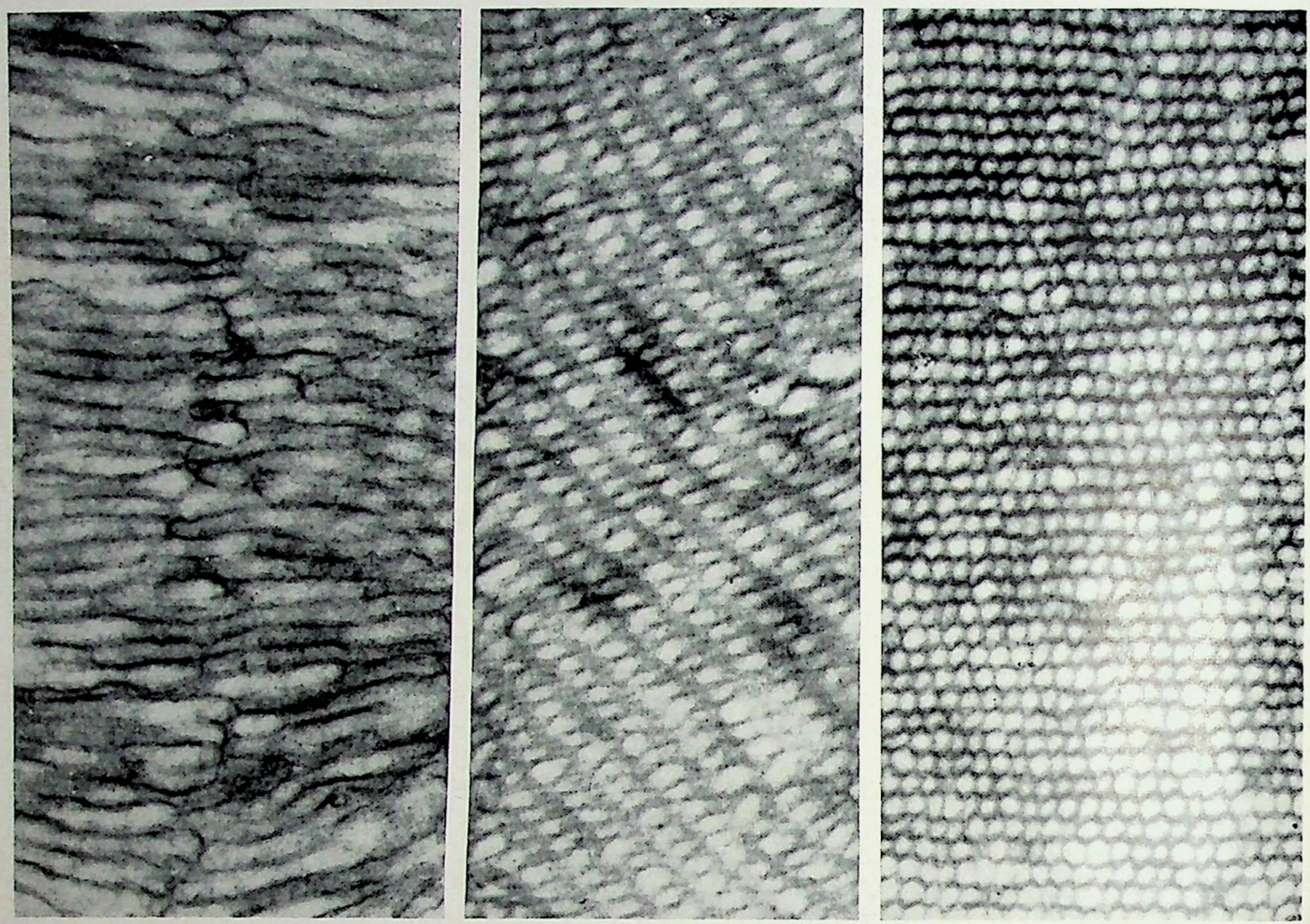


FIG. 1. Electron micrographs of rhabdom of an ommatidium of an arthropod compound eye (*Limulus*), showing honeycomb-like arrangement of osmium-staining membranes. Left: plane of section perpendicular to optical axis of ommatidium. Center: oblique section. Right: section in an axial plane. Height of figure approx 2 μ. Courtesy W. H. Miller.

receptor properties, such as the loss of sensitivity in the light and its recovery during dark adaptation, can be explained qualitatively by these elementary biochemical processes in the visual receptors. Moreover, a simple model of the photochemical system of the receptor was used by Hecht to explain quantitatively many psychophysical measurements of vision.[7] His formulations remain the most comprehensive and successful theoretical treatment of visual-receptor physiology, although his model is oversimplified and the theory needs reworking in light of recent developments in biochemistry and physiology.

The primary photosensitive pigment of the visual receptor is present in a structured system. Electron-microscope studies show a profusion of osmium-staining membranes in visual receptors. Sjöstrand[8] has shown that the outer segments of the receptors of the vertebrate retina have a lamellar structure. The rod outer segment is thus a stack of thin plates crowded with rhodopsin. In the arthropods, instead of a lamellar system, the part of the receptor cell (the rhabdom) that presumably contains the visual pigment is composed of myriads of microvilli densely packed to form a honeycomb-like structure. Figure 1 is an electron micrograph of the rhabdom of an ommatidium of an arthropod compound eye.[9] In the vertebrates, the outer segments of the retinal rods and cones have been shown to be derivatives of cilia.[10,11] In the arthropods, where cilia are extremely rare, there is no evidence of a ciliary derivation. In some mollusks, however, there is a different system of membranes and again the structures are derived from cilia.[12] Exactly how the visual pigment is arranged within any of these membranous structures is not known, though there have been speculations on this point.

Visual receptors have evolved into light detectors that are so sensitive that they work at the limit set by the quantum nature of light. A human observer is able to see a flash of light that contains only about 100 quanta, measured at the cornea of the eye. After correction for losses in transmission through the ocular media and failure of the visual purple to be present in sufficient amount in the retina to absorb all of the quanta that fall on it, this figure comes down to something of the order of 10 quanta.[13] This aspect of visual physiology has been extensively studied and is well reviewed in a recent article by Pirenne.[14] Obviously, it is of great significance to visual performance, especially at low

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

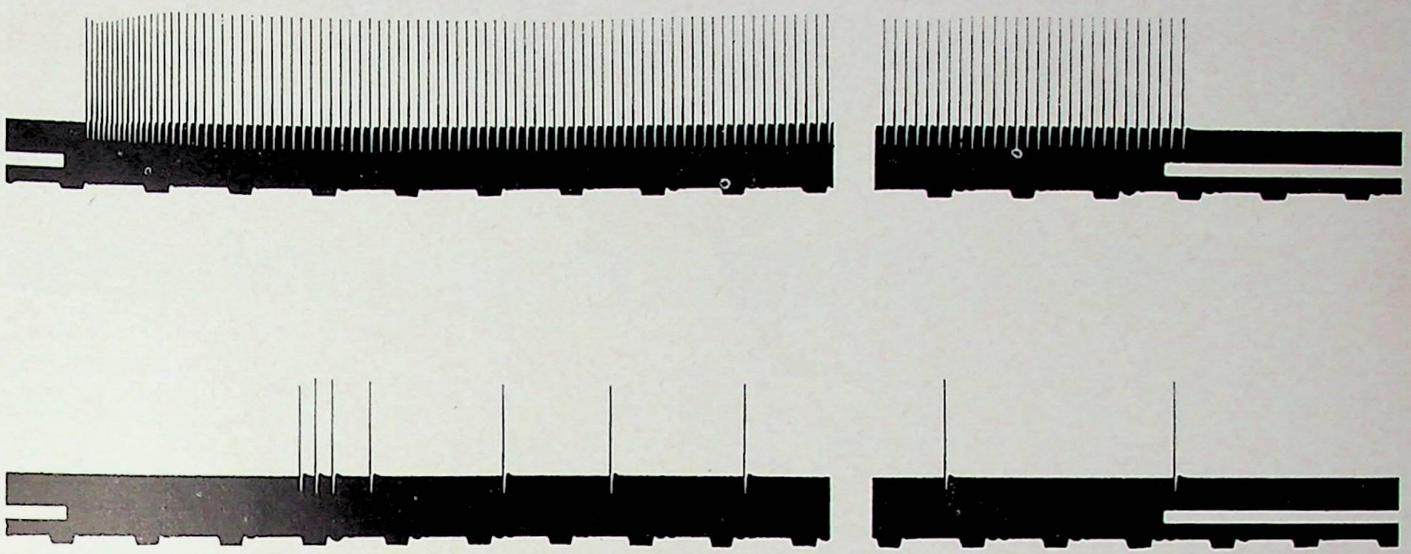


FIG. 2. Oscillograms of the electrical activity in a single optic-nerve fiber (eye of *Limulus*) in response to prolonged steady illumination of the facet of the eye innervated by the fiber.[15] Each spike is the "action potential" associated with the passage of a nerve impulse in the fiber. For the top record, the intensity of stimulating light was 10^1 times that used for the bottom record. Signal of exposure to light blackens out the white line above the time marks. Time marked in $\frac{1}{5}$ sec.

illuminations. In the short "action time" of the retina, so few quanta are needed that the retinal image, though visible, is too "grainy" to be seen with high resolution. Also, at threshold, seeing is uncertain. Indeed, the statistical uncertainty at threshold can be explained almost entirely by the fact that a very few quanta suffice to excite a response. Nevertheless, the visual threshold is sharp enough so that it is quite certain that a human observer cannot see just one quantum, although exactly how many are needed is still a matter of controversy. The small amount of light that is just visible can be seen if it is spread over a retinal area containing about five hundred rods. This must mean that near threshold there is almost no chance of any one rod receiving more than one quantum, and that the cooperative activity of several rods is necessary to reach the threshold of vision. Thus, a single quantum of light absorbed within the stack of plates comprising the outer segment of a rod is sufficient to excite that rod, causing it to transmit some kind of nervous influence that can sum with similar influences from several other rods to reach the threshold for a behavioral response. In this retinal summation, one has an example of the simplest kind of neural integrative action, exerted at the very threshold of vision.

The end result of receptor excitation is the generation of nervous influences in its attached nerve fiber. It has not yet been possible to record the neural activity of the receptors (rods and cones) of the vertebrate retina, but some invertebrate eyes afford an opportunity to record optic activity that appears to be very close to the action of the primary receptors. The eye of the common horseshoe crab, *Limulus*, is particularly favorable for the study of the action of single receptor units. This eye is a coarsely facetted compound eye. Individual receptor units corresponding to each facet (ommatidia) can be separately illuminated, and the electrical activity of the optic nerve fiber from such a unit can be recorded.[15]

The neural activity recorded from one of the receptor units of the eye of *Limulus* consists of trains of uniform nerve impulses similar in all respects to the sensory discharges observed in nerve fibers in all of the higher animal forms (Fig. 2). As in all receptors, the higher the intensity of the stimulus, the higher the frequency of the impulses with which the receptor responds. In the visual receptor, it is noteworthy that frequency changes over a relatively small range for a large range of light intensity: the dynamic range of a single receptor is five or six orders of magnitude. Roughly, the relation between frequency of discharge and intensity of light is a logarithmic one (Fechner's Law). Thus, the transducer mechanism of the visual receptor covers a large range and is adapted to signal the ratios of stimulus values. In our own visual experience, values of light and shade stay more or less fixed, no matter what the ambient level of illumination, over a large range. In such situations, stimulus ratios stay constant and the visual receptors yield approximately a fixed difference in the frequency for a given ratio of stimulus values, even though the absolute differences may vary widely.

Another important receptor property is illustrated in Fig. 2. The discharge of nerve impulses begins at a high frequency when the light is turned on, but the frequency of the discharge subsides in a fraction of a second to a considerably lower level, which is then maintained with only slight diminution as long as light continues to shine on the receptor. This sensory adaptation is manifested by all other receptors, some to a far greater extent than others. As a result of sensory adaptation, receptors provide a somewhat distorted report of the stimulus events, such as to accentuate any sudden change. Sensory trans-

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

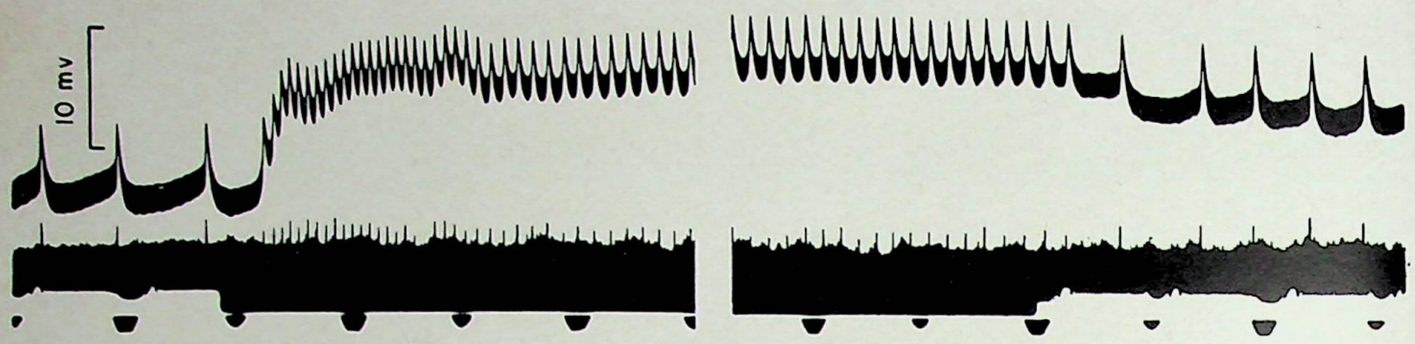


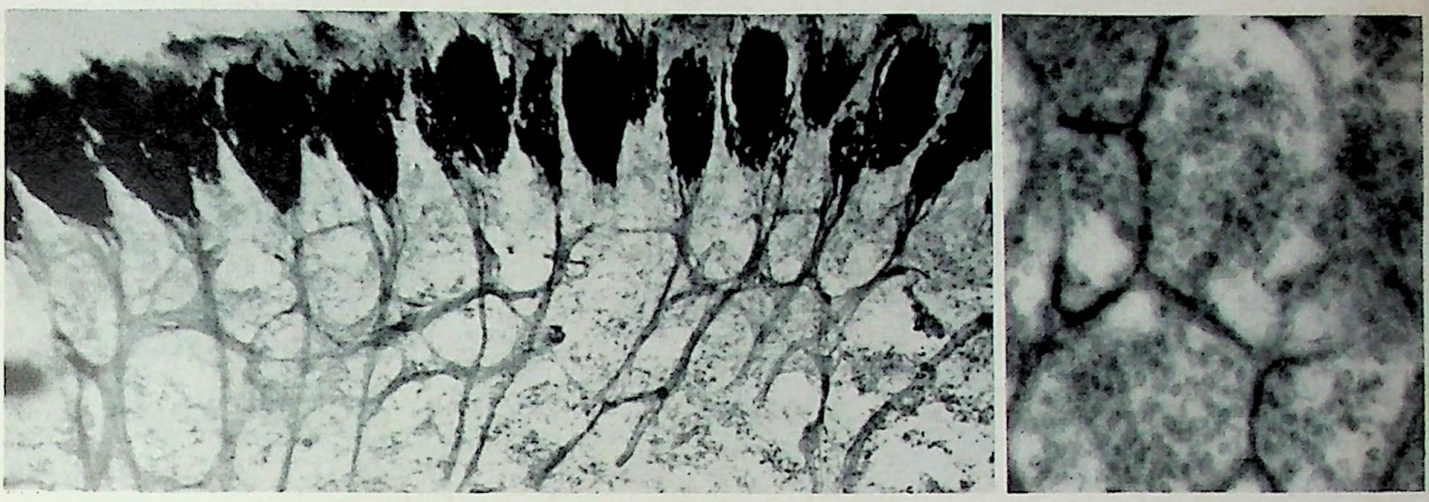
FIG. 3. Electrical responses to illumination of a single receptor unit (ommatidium) in the eye of *Limulus*.[15] Recorded (upper trace) by a micropipette electrode (tip diam 1 μ) in the neuron of the ommatidium, and simultaneously (black edge just below upper trace) by a pair of wire electrodes over which was slung the nerve bundle from the ommatidium containing the neuron's axon. At the beginning of the record, the microelectrode base line records the resting level of the electrical polarization of the cell membrane, at a potential about 50 mv negative to the solution bathing the outside of the cell. When the ommatidium was illuminated (black band above the time marks), there was a partial depolarization of the neuron (potential becoming less negative: rise in the base line) accompanied by an increase in the frequency of the spike-like deflections, each one of which was synchronous with the discharge of an impulse in the nerve bundle (small spikes on the black edge). Time marked in $\frac{1}{5}$ sec.

ducers are not concerned so much with a faithful representation of the world as with a useful one, and it is especially useful to the organism to accentuate the *changes* that occur in external conditions. If the illumination on a visual receptor unit is given a small increment, the receptor response consists of a modulation of the discharge of impulses in which there is an exaggeration of frequency changes at the onset and again when the increment is turned off. This permits the receptor to signal small changes, and still possess a large dynamic range. But what may be even more important, the suddenness of the changes enhances their stimulating effectiveness. Thus, the inherent properties of the sensory receptors determine how the patterns of neural activity they generate will represent the stimulus events. This is a first step in the processing of information for use by the organism.

As yet not much is known about the nature of the excitatory processes following the initial photochemical reaction in visual receptors until one comes near the end of the receptor process. In the eye of *Limulus*, it has been possible to learn a little about the actual production of nerve impulses in the axon of the excited neuron in the receptor unit. By the use of a micropipette electrode penetrating the sensory structure of the ommatidium, changes in the electrical polarization of the cell membrane of the neuron in the ommatidium have been recorded.[15–17] These changes are associated with the trains of impulses initiated by this cell when the receptor unit is illuminated (Fig. 3). When light is turned on, the cell membrane becomes somewhat depolarized, and simultaneously there is a speeding up of the discharge of impulses in its axon. Such depolarization is referred to as a "generator potential,"[18] in the belief that the nerve impulses are generated by local electric currents flowing as a result of the difference in potential between the axon and the depolarized cell body (or more probably, in the present case, the depolarized dendritic process of the cell which penetrates the rhabdom of the ommatidium). The degree of depolarization depends on the intensity of the stimulating light and in turn determines the frequency of the relaxation oscillations of the membrane of the initial segment of axon in the region where it leaves the cell body and from which the propagated impulses take off. The discharge of trains of impulses by depolarized neurons is a familar process in neurophysiology. For the photoreceptor, the question is how the initial photochemical reaction produces the depolarization and the ensuing "generator potential." About this, almost nothing is known.

As discussed in the foregoing, the receptor itself by its inherent properties does a certain amount of processing of the information from the outside world. It is concerned with the report only of certain aspects of the physical stimulus that acts on it, and it is not necessarily a high-fidelity recording device. Built as it is, it selects certain features of the stimulus pattern for accentuation. The next step in the processing of sensory information in the visual system concerns the distribution of light over the entire population of visual receptors. A retina, whether in a vertebrate or an arthropod, is more than a mosaic of independent detecting elements. In the vertebrates, it is well known that the retina is a highly organized nervous center. It is really a part of the brain closely applied to a mosaic of sensory receptors. The first step in the neural analysis of the pattern of the retinal image requires the intercomparison of what happens in the various differently stimulated receptors, and a modification of the pattern of neural activity to accentuate important features of the spatial distribution of light over the receptor mosaic. Evidently, it is profitable to do this close to the point where the information is being picked up. In the vertebrate retina, the early neurons in the visual pathway are spread out in correspondence with their associated receptors, and many of the processes in the first step of neural integration apparaently can be done most effectively in the retina itself. This is not a simple process; patterns of activity observed in the optic-nerve fibers in the verte-

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA



(a) (b)

FIG. 4. The plexus of the compound eye of *Limulus*. (a) Light micrograph of a section cut through the eye in a plane perpendicular to its external surface (cornea removed), showing on its upper border a row of the heavily pigmented ommatidia, from each of which emerges a small bundle of nerve fibers (stained with silver by Samuel's method) that contains, together with small fibers, the axon of ommatidium neuron. Connecting these bundles are festoons of fibers, with clumps of neuropile that appear at this magnification as condensations in the meshes of the plexus. Width of figure = 2.2 mm. Photograph by W. H. Miller [from H. K. Hartline, H. G. Wagner, and F. Ratliff, J. Gen. Physiol. **39**, 651 (1956)]. (b) Electron micrograph of a portion of a clump of neuropile in the plexus, showing a few outlines of the fibers composing the clump, within which are numerous small circular outlines interpreted as synaptic vesicles. Width of figure = 1.2 μ. Photograph by W. H. Miller.[19]

brate retina are very complex, and their analysis is difficult. In simpler visual systems, integrative processes can be more readily analyzed. The eye of *Limulus* again affords a good oportunity for such studies.

The neural structure of the compound eye of *Limulus* is much simpler than that of the vertebrate retina or the eyes of more highly developed arthropods, but it is nevertheless a retina: the units of the receptor mosaic are interconnected by a network of nerve fibers [Fig. 4(a)]. The nerve fibers from the ommatidia branch profusely on their way out of the eye to form the optic nerve. Festoons of these branches connect each receptor unit with its neighbors. There are no nerve-cell bodies in this plexus of interconnections, as in more complex retinas, but there are numerous knots composed of a felt-work of very fine branchlets closely intertwined. The fibers in these clumps of "neuropile" are packed with "vesicles" typically present in synapses [Fig. 4(b)]. Evidently, the clumps of neuropile are synaptic regions, where influences are transmitted from one set of

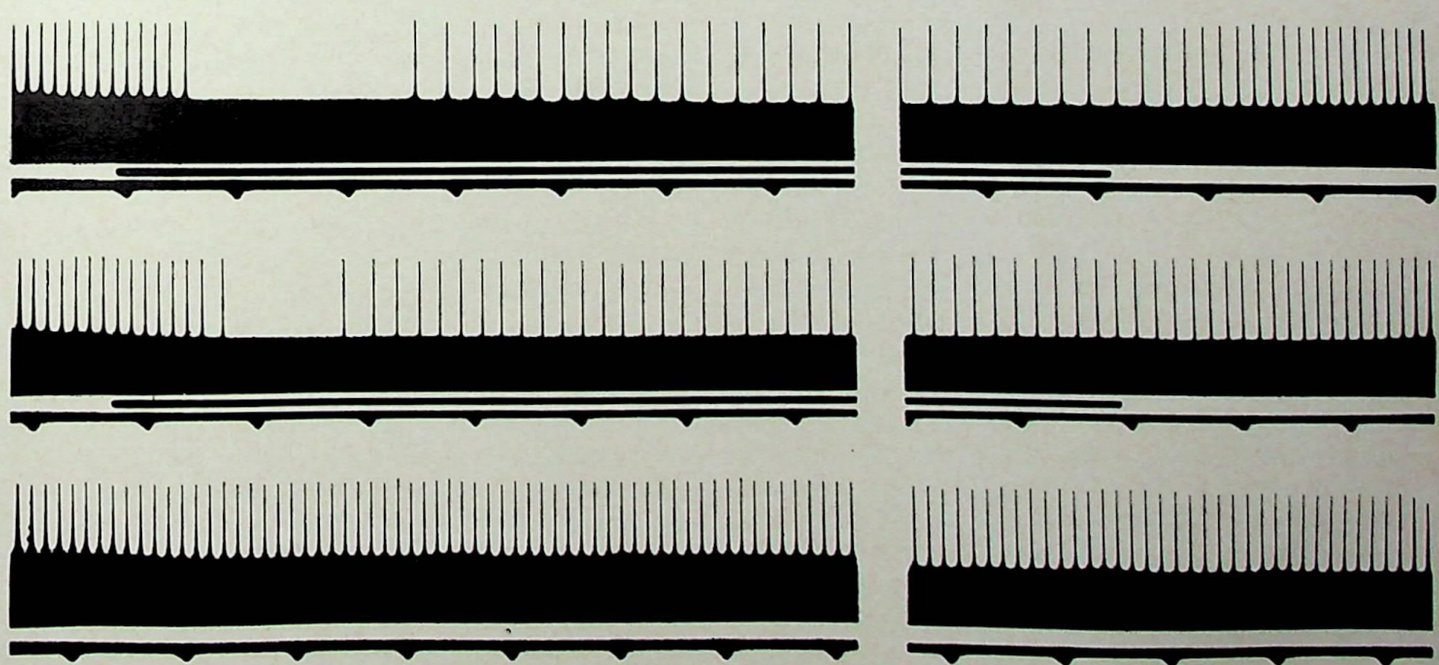


FIG. 5. Oscillograms of nerve action potentials, showing inhibition of the impulses in a single optic-nerve fiber of *Limulus*. The ommatidium of the eye from which the fiber arose was illuminated steadily at a fixed intensity, beginning 3 sec before the start of each of the records; adjacent ommatidium were illuminated during the interval signalled by the blackening out of the white line above the time marks, in the upper two records. In the top record, the intensity of the illumination of adjacent receptors was ten times that used in the middle record. Bottom record is a control (no adjacent illumination). Time in $\frac{1}{3}$ sec. Experimental arrangement as in Fig. 6(a).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

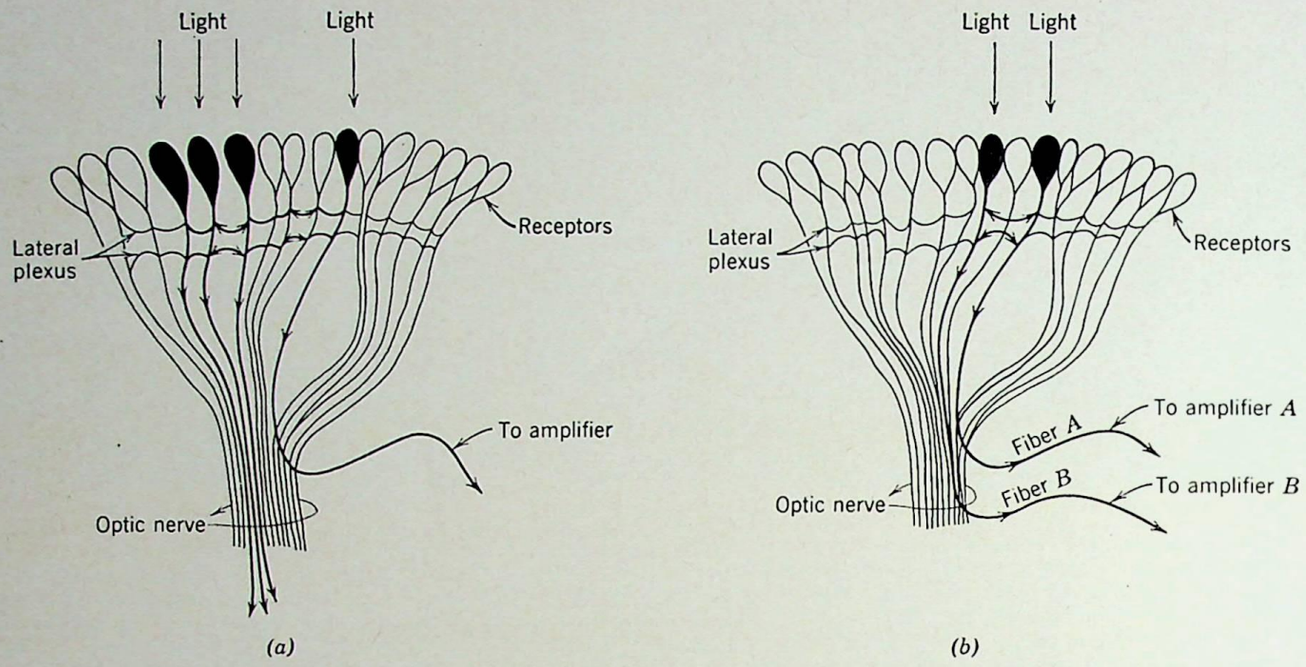


FIG. 6. Schematic diagrams of experimental arrangements used. (a) Experiment of Fig. 5 (inhibition of test ommatidium by illumination of nearby ommatidia). (b) Experiments of Figs. 7 and 8 (mutual inhibition of two ommatidia that were close to one another).

branches to another.[19] Based on this structural organization is a simple functional organization: each ommatidium tends to inhibit the activity of its neighbors. This influence is indeed exerted over the plexus of nerve fibers, for, by cutting the interconnecting branches to an ommatidium, the influence of its neighbors on it can be abolished.

The inhibition that is exerted on an ommatidium by its neighbors is illustrated in Fig. 5. In the experiment from which these records were taken, the discharge of impulses was recorded in a single optic nerve fiber in response to illumination, by a small spot of light, of the facet of the ommatidium from which that fiber arose [Fig. 6(a)]. During steady illumination of that one ommatidium alone, a steady discharge of impulses resulted (bottom record). When, during steady and continuous illumination of this "test" ommatidium, light was caused to shine also on other ommatidia in neighboring regions of the eye (top and middle records of Fig. 5), the frequency at which the test ommatidium discharged impulses was reduced.

Figure 5 also shows that strong illumination of the adjacent region produced a greater depression of frequency than weak illumination. It has also been shown that the magnitude of the inhibition exerted on an ommatidium is greater the larger the number of neighboring ommatidium that are stimulated. Thus, the inhibitory influences from many neighbors can combine to increase the net effect they produce. Also, the inhibition exerted on an ommatidium by its neighbors is greater the closer they are to it. Ommatidia that are separated by a distance exceeding 4 or 5 mm have no effect on one another.

Inhibition in the eye of Limulus is exerted mutually by the receptor units.[20] Each ommatidium, being a neighbor of its neighbors, inhibits them as well as being inhibited by them. This is shown in Fig. 7, obtained by recording activity simultaneously in the optic nerve fibers from two independently illuminated ommatidia, close to each other in the eye [Fig. 6(b)]. The frequency of each receptor unit was lower when both were illuminated together than when each was illuminated by itself. When this experiment is performed using various intensities on the two receptors in various combinations, it has been shown that the amount by which the steady frequency of discharge of each receptor unit is lowered depends on the degree of concurrent activity in the other, and is indeed a linear function of the frequency of its discharge (Fig. 8). Thus, the response of one receptor unit is determined by the excitation furnished by the stimulating light shining on it, diminished by the inhibitory influence from the second receptor, which in turn depends on the resultant of the excitation furnished by its own stimulus and the inhibition exerted on it by the first. This mutual interdependence of any two neighboring receptor units may be described by a pair of simultaneous equations, linear in the frequencies of the discharges.

When more than two interacting receptors are illuminated simultaneously, each is subject to the combined inhibitory influences from all of the others. The law that determines how the inhibitory influences from several active receptor units combine in affecting the activity of a neighboring unit has been found by experiment: if the influences on a given unit are measured by the reduction they produce in its frequency of nerve impulse discharge, the combined effect of all of the other units is simply given by the sum of the influences

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

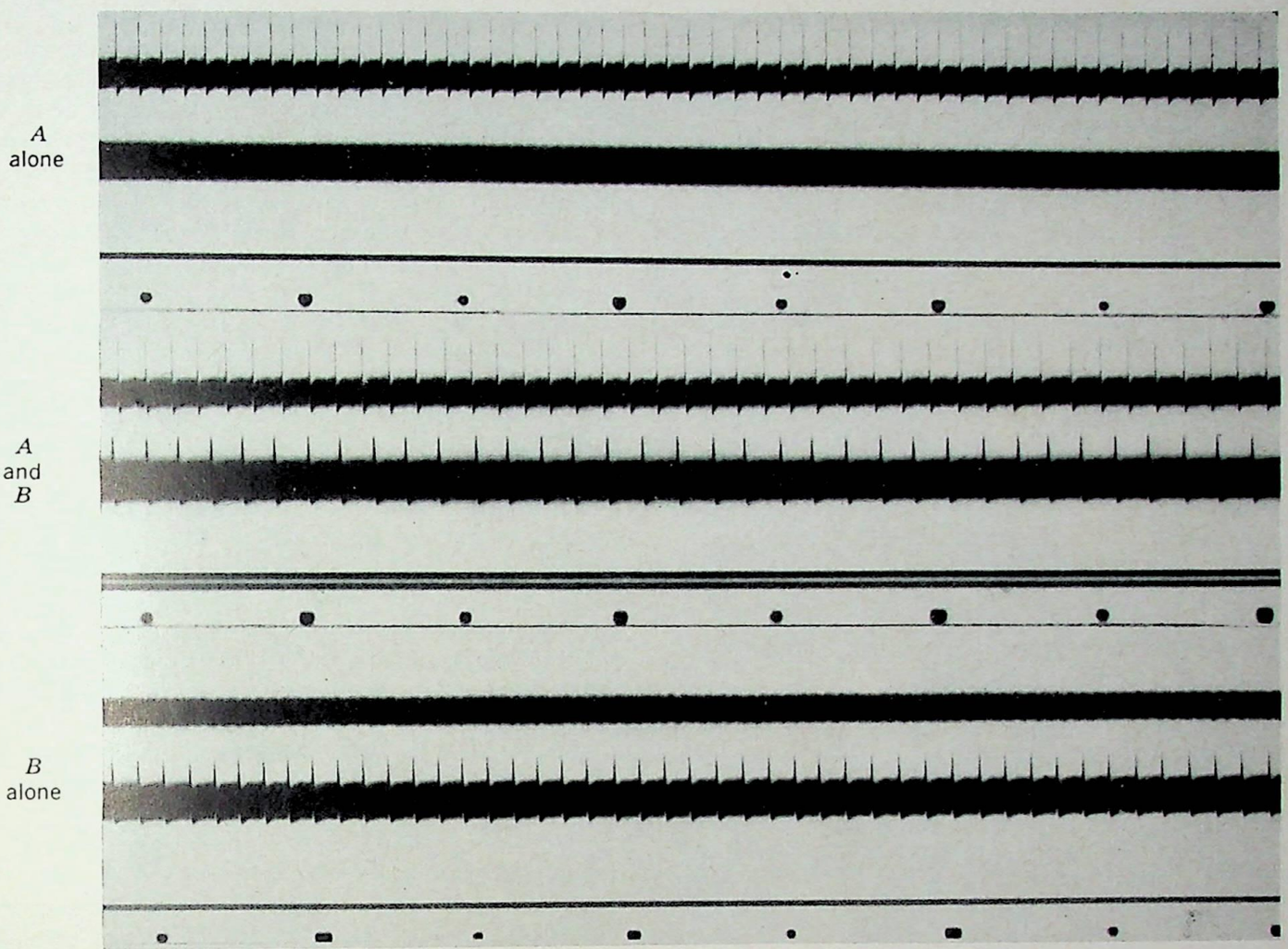


FIG. 7. Mutual inhibition of two ommatidia close to one another in the eye of *Limulus*, steadily illuminated at fixed intensity on each. Experimental arrangement as in Fig. 6(b). Time (black dots) in $\frac{1}{5}$ sec.

exerted by each.[21] The responses of a set of n interacting receptor units, measured by the frequencies of their optic nerve discharges, are therefore expressed by a set of n simultaneous linear equations,

$$r_p = e_p - \sum_{j=1}^{n} K_{pj}(r_j - r_{pj}^0) \quad p = 1, 2 \cdots n.$$

In these equations, r_p stands for the response of the pth unit (measured by its steady frequency of impulse discharge) when it is illuminated steadily together with the other units. Its excitation, e_p, is measured by the response it has when it is illuminated alone. Each constant K_{pj} is the coefficient of the inhibitory action of the jth receptor on the pth (usually less than 0.2) and each constant r_{pj}^0 is the threshold of that action. Terms for which $j = p$ are usually omitted. The equations as written apply only to those units and that range of activity for which r_j is not less than r_{pj}^0. As a rule, the closer the interacting elements are to one another, the larger the K's and the smaller the r^0's. Exceptions are often found, however, and it is not yet possible to state the statistical law governing the effects of distance on the inhibitory interaction.

If N small groups of receptors are considered, each group uniformly illuminated and assumed to consist of receptors with similar properties exerting equal actions, the foregoing set of equations may be reduced to N simultaneous equations with lumped coefficients representing the group interactions. Applied to three interacting receptors or receptor groups, the theory outlined in the foregoing can account quantitatively for a number of effects that have been observed with various experimental configurations of retinal illumination. Thus, if a test receptor is located midway between two groups of receptors that are themselves too far apart to interact, the combined inhibitory action of these two on the test receptor is equal to the sum of the separate actions of each, unless the test receptor itself has an appreciable effect upon them. If the two groups are close together and both are near the test receptor, their combined inhibitory effect will be less than the sum of their separate effects since they inhibit one another mutually when both are active. If a group of receptors is too far from a test receptor to influence it directly, it may nevertheless affect its response by inhibiting a second group located close to the test receptor, thereby releas-

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

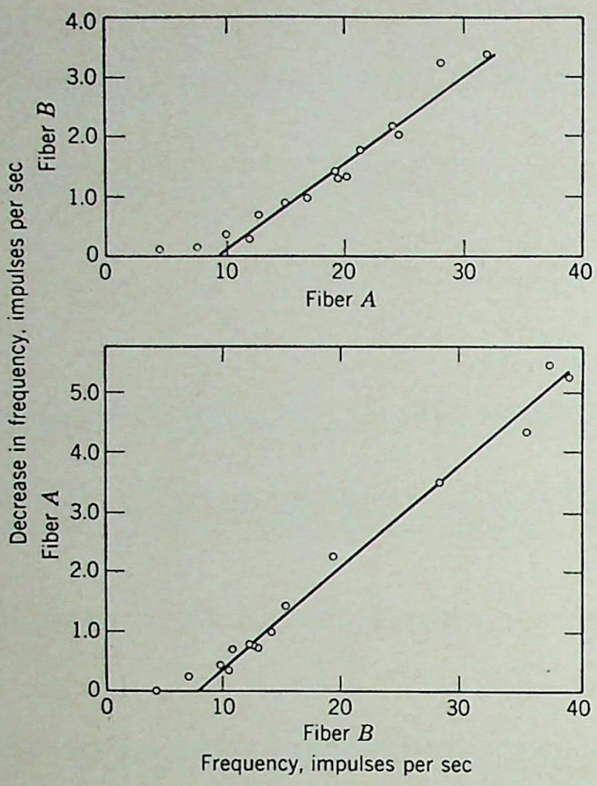


FIG. 8. Mutual inhibition of two ommatidia close to one another in the eye of *Limulus*, illuminated independently at various levels of intensity in various combinations. Amount of inhibition (decrease in frequency) of each ommatidium plotted as a function of concurrent level of response (frequency) of the other. From an experiment similar to that of Fig. 7, experimental arrangement as in Fig. 6(b).

ing the test receptor from the inhibition exerted by the second group. Such "disinhibition" illustrates how indirect effects may be exerted beyond the limits of direct influence and, in principle at least, extend over the entire mosaic of interdependent receptor units.

The inhibitory interaction just described may be considered a simple integrative mechanism that takes place at or close to the level of the receptors themselves. Because of it, patterns of optic nerve fiber activity yield a distorted representation of the patterns of incident illumination. This distortion, however, serves a useful function, for it is clear that inhibitory interaction must enhance contrast: brightly lighted elements in the receptor mosaic inhibit the dimly lighted ones more than the latter inhibit the former. If, as in the eye of *Limulus*, mutual inhibition is greater between close neighbors than distant ones, contrast will be greatest near regions of steep intensity gradients and borders and edges in the retinal image will be "crispened." Phenomena of border contrast are illustrated in our own vision by the light and dark bands bordering a penumbra (Mach bands), and by the fluted appearance of an optical step-wedge or of shadows cast by multiple light sources. Inhibitory interaction is probably one of the mechanisms in our own visual systems that gives rise to these phenomena.[22]

Direct experimental demonstration of the "crispening" of the contours by inhibitory interaction can be made, using the eye of *Limulus*.[23] The discharge of impulses is recorded from a "test" receptor near the center of the eye as the eye is caused to scan slowly a pattern of illumination containing a gradient of intensity. When all of the receptors are masked except for the one from which activity is being recorded, a faithful representation is obtained of the distribution of intensity in the image viewed. But when the mask is removed, so that all of the receptors view the pattern, maxima and minima in the frequency of the test receptor discharge occur, corresponding to the regions bordering the gradient. These resemble in form and location the "Mach bands" seen by a human observer viewing the same pattern.

Interaction is known to take place in other sense organs. In the ear, von Békésy[24] has suggested that inhibitory interaction may be important in increasing pitch discrimination. Indeed, Galambos and Davis[25] have demonstrated inhibition of the activity of single auditory nerve fibers by tones differing in frequency from those used to excite the fibers. Also, von Békésy[26] has demonstrated "contrast" effects in the tactile stimulation of the skin, suggesting inhibitory interaction over considerable distances over the surface of the body.

In the higher nervous centers, integrative processes of great complexity take place. In the retina of the vertebrate eye, which—even though located in the peripheral sense organ itself—is nevertheless a nervous center of high order, there is an intricate interplay of excitatory and inhibitory interactions.[27] As a result, diverse and labile patterns of optical nerve fiber activity are generated.[28,29] In the vertebrate retina, to a much greater degree than in the primitive retina of *Limulus*, the patterns of afferent nervous activity are greatly modified to accentuate significant features of information about the environment. The process of neural integration is well begun by the time the afferent messages are transmitted to still higher centers in the brain.

BIBLIOGRAPHY

[1] H. J. A. Dartnall, *The Visual Pigments* (Methuen and Company, Ltd., London, 1957).

[2] G. Wald and P. K. Brown, Science **127**, 222 (1958).

[3] G. Wald, P. K. Brown, and P. H. Smith, J. Gen. Physiol. **38**, 623 (1955).

[4] G. Wald, Am. Scientist **42**, 73 (1954).

[5] R. Hubbard and A. Kropf, Proc. Natl. Acad. Sci. U. S. **44**, 130 (1958).

[6] F. W. Campbell and W. A. H. Rushton, J. Physiol. **130**, 131 (1955).

[7] S. Hecht, Physiol. Rev. **17**, 239 (1937).

[8] F. S. Sjöstrand, J. Cellular Comp. Physiol **42**, 15 (1953).

[9] W. H. Miller, J. Biophys. Biochem. Cytol **3**, 421 (1957).

[10] E. DeRobertis, J. Biophys. Biochem. Cytol. **2**, 209 (1956).

[11] K. R. Porter in *Harvey Lectures* (Academic Press, Inc., New York, 1957), Vol. LI, p. 175.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

[12] W. H. Miller, J. Biophys. Biochem. Cytol. 4, 227 (1958).
[13] S. Hecht, S. Schlaer, and M. H. Pirenne, J. Gen. Physiol. 25, 819 (1942).
[14] M. H. Pirenne, Biol. Revs. Cambridge Phil. Soc. 51, 194 (1956).
[15] H. K. Hartline, H. G. Wagner, and E. F. MacNichol, Cold Spring Harbor Symposia Quant. Biol. 17, 125 (1952).
[16] E. F. MacNichol, Am. Inst. Biol. Sci. 1, 34 (1956).
[17] T. Tomita, Japan. J. Physiol. 6, 327 (1956).
[18] R. Granit, *Sensory Mechanisms of the Retina* (Oxford University Press, London, 1947).
[19] F. Ratliff, W. H. Miller, and H. K. Hartline, Ann. N. Y. Acad. Sci. 74 (2), 210 (1958).
[20] H. K. Hartline and F. Ratliff, J. Gen. Physiol. 40, 357 (1957).
[21] H. K. Hartline and F. Ratliff, J. Gen. Physiol. 41, 1049 (1958).
[22] G. A. Fry, Am. J. Optom. and Arch. Am. Acad. Optom. 25, 162 (1948).
[23] F. Ratliff and H. K. Hartline, J. Gen. Physiol. (to be published).
[24] G. von Békésy, Physik. Z. 19, 793 (1928).
[25] R. Galambos and H. Davis, J. Neurophysiol. 1, 287 (1944).
[26] G. von Békésy, J. Acoust. Soc. Am. 30, 399 (1958).
[27] R. Granit, *Receptors and Sensory Perception* (Yale University Press, New Haven, 1955).
[28] H. K. Hartline in *Harvey Lectures* (Academic Press, Inc., New York, 1942), Vol. XXXVII, p. 39.
[29] S. W. Kuffler, J. Neurophysiol. 6, 37 (1953).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

56

Mechanisms of Synaptic Transmission

BERNARD KATZ

Department of Biophysics, University College, London, W. C. 1, England

INTRODUCTION

THE propagation of an impulse along a nerve or muscle fiber is brought about by two coupled processes: (i) cable transmission, which allows an electric potential change to spread along a short distance, but with rapid attenuation, and (ii) a boosting mechanism by which the full signal strength is regenerated at each point. If either of these processes is interfered with, the signal will be blocked and will fade out locally. The cable mechanism depends upon the continuity of the fiber structure, with a relatively low-resistance core and high-impedance surface layer. During the impulse, sufficient current must be able to flow forward along the inside of the axon and outward through the resting membrane to stimulate it. If one were to close the core with a high-resistant transverse membrane, or to place a low-resistance shunt across the fiber surface, transmission would be impaired and probably would fail at that point. The very reason why thousands of axons packed together within one nerve bundle can conduct their messages independently, without mutual interference, rests on the absence of structural continuity, and lo of an effective cable connection between them.

The purpose of this paper is to consider what happens at "synapses," the points of contact between one nerve cell and the next, or between nerve and muscle fiber. There is no sign of cytoplasmic continuity between the different cell units. Electron-microscope evidence shows that the membranes of the synapsing cells are arranged in close proximity, though in general they do not seem to fuse or to come into intimate contact. The electron micrographs, however, do not reveal anything about the electrical properties of the contacting surfaces; and one has no means of guessing intuitively whether or not an effective cable linkage exists across the synapse.

ELECTRICAL AND CHEMICAL TRANSMISSION

There are, in principle, two basically different modes of synaptic transmission, electrical and chemical. *Electric* transmission implies that in spite of apparent morphological complexities, an effective local-circuit connection exists which allows sufficient current to pass from one cell to the next to restimulate it. In other words, the transmission is one continuous process without any essential change at the synapse. There must, of course, be *some* difference; for all synapses have the property of functioning only in one direction unlike nerve fibers which can conduct impulses with equal facility in either direction despite the fact that in normal life, because of their terminal synaptic connections, they are used for one-way traffic only.

Chemical transmission implies the intervention of an entirely different process specific for the synaptic area. It presupposes that the ordinary cable-connection is interrupted at the contact point between the cells, and replaced by the agency of a chemical mediator. A specific chemical stimulant which is synthetized and stored inside the nerve terminal is liberated by the nerve impulse. When the substance is released, it attaches itself to special receptor molecules in the surface of the contacting postsynaptic (or effector) cell. This chemical combination leads to a membrane change which gives rise to a local depolarization of the effector cell. When the depolarization exceeds the threshold level, a new action potential is set up which then travels along the cell in the manner previously described. Thus one has, interposed between two separate waves of propagated electric activity, a secretion of a specific substance from one cell, and a chemoreceptor reaction in the surface of the next cell.

Now, it is not possible to predict without thorough experimental examination which of these two modes of transmission occurs at a particular type of synapse. Various attempts have been made to generalize by drawing analogies from the few cases which have been explored; and in recent years the view has become prevalent that transmission is llkely to be chemical at all synapses, but that a variety of substances are being employed in different cases, and only a few of them, like acetycholine and noradrenaline, have so far been identified. This view, however, seems a little too sweeping, and the discussion is begun, therefore, by quoting an example of electric transmission which has just been brought to light by the work of Furshpan and Potter.[1] This may well be an exception to the rule, but it provides a definite warning against too much generalization in this field.

Furshpan and Potter used a "giant" synapse in an abdominal ganglion of the crayfish cord. This is a contact point between a very large nerve fiber which runs through the central nervous system of the crayfish and a somewhat smaller motor axon which emerges from the cord to supply the flexor muscles of the "tail." This synapse was chosen for two reasons: (a) because the large size of the two contacting cells made it possible for a pair of microelectrodes (one to pass current and one to measure membrane potential) to be introduced into each of them; (b) it seemed *a priori* that electric transmission might be feasible at this synapse, more so than at other types where minute nerve endings usually

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

terminate in contact with a huge cell (which implies, from the electrical point of view, a poor mismatch, very little current being available from the high-impedance terminals to discharge the large surface of the postsynaptic cell). Furshpan and Potter were able to show that the membrane contact of this special synapse acts as a good electric rectifier, allowing current to pass relatively easily from the presynaptic to the postsynaptic cell, but not in the reverse direction. In other words, at this particular synapse, an adequate cable connection exists between the interior of the two cells, in the *normal* direction of impulse travel alone. Provided the internal potential of the prefiber was *higher* than that of the postfiber, electric current could flow across and influence the membrane potential of the adjoining cell. The result is that a depolarization such as occurs during the impulse can be transmitted in the normal *orthodromic* direction, but not the other way. And, conversely, a local hyperpolarization, produced experimentally by passing a current inward through the fiber membrane was found to be transmitted only in the *antidromic* direction (from postfiber to prefiber) but not the other way. Thus, at the giant synapse of the crayfish, one has a case—so far, the only known example—of electric transmission, in which the action current generated by the arrival of an impulse in the presynaptic cell is passed on without finite delay and can directly depolarize and thereby excite the postfiber. One-way transmission is owing to the valve-like one-way resistance of the synaptic contact membranes.

There are only a few giant synapses available in nature allowing such a direct experimental approach to both sides of the junctional region. In most cases, the presynaptic nerve endings are too small to be tackled with intracellular electrodes and their electrical behavior has to be inferred from a more indirect approach. It is of great interest, however, that at another giant synapse, in the stellate ganglion of the squid, Bullock and Tasaki and Hagiwara[2,3] obtained evidence of a different kind; they observed a definite local delay in the propagation of the electrical change, indicating a stoppage of the local-circuit transmission at the junction; there was no detectable transfer of subthreshold cable signals in either direction. These observations provide another fair warning against attempts to generalize about synaptic mechanisms.

If one now takes an entirely different case—namely, the skeletal neuromuscular junction—one finds here one of the few examples of a synapse where chemical transmission has been firmly established. It was shown by Dale and his colleagues that a specific cholinester, almost certainly identical with acetylcholine, is released from the active motor-nerve endings. This substance is a very potent local stimulant and, provided it is applied rapidly to the junctional end-plate region of a muscle fiber, causes a local depolarization of the cell membrane and sets up propagated impulses and contraction in the fiber. The chemical effect is localized to the synaptic area of the muscle surface; it is at these points exclusively that a number of chemical blocking agents, like curare, act (apparently by a competitive attachment to the acetylcholine receptors). By histochemical methods, a high local concentration of a specific enzyme, acetylcholinesterase, has been found at the same point; the apparent purpose is to hydrolyze the transmitter substance within a very short time after it has exerted its action. Electrical studies have shown that there is an irreducible delay of 0.5 to 1 msec between the arriving nerve impulse and the start of the so-called end-plate potential (which is the local depolarization in the muscle fiber produced by the transmitter substance). There is no cable transfer of electric current, of either polarity, directly from the nerve axon to the muscle fiber. When potential changes are imposed on the terminal portion of the motor nerve, these changes do not spread beyond the nerve terminal, but can be shown to increase the rate at which acetylcholine is being released from the nerve endings and so, indirectly, to influence the membrane potential of the muscle fiber.

The most direct way of establishing chemical transmission by nerve impulses would be to show that a substance is released, on nerve stimulation, into the circulating fluid, and when applied to a remote effector cell produces the same, excitatory or inhibitory, action. This has been achieved in a few cases, notably the classical experiment of Otto Loewi in which he demonstrated the role of acetylcholine as the transmitter of nervous inhibition of the heart beat. Usually, such a direct demonstration is not feasible because of the enormous dilution of the transmitter agent, on the way from its primary point of release and action to the assaying tissue. Indeed, this discrepancy between (i) the amount of acetylcholine (ACh) which has to be applied artificially to stimulate a muscle, and (ii) the much smaller quantities which are released into the perfusion fluid from stimulated nerve endings, has been used as an argument against the validity of the chemical-transmitter theory. By a method of microscopic ionophoresis, much more effective applications have been made recently, and it has been possible to show that as little as 10^{-16} g equiv of ACh can give rise to an effective, superthreshold, depolarization of the muscle fiber. This is still a few hundred times more than the amount per impulse recovered in the earlier perfusion experiments, but it must be remembered that even with the best micropipettes one cannot reduce the average diffusion distance to that of the natural synaptic contact. The remaining quantitative discrepancy is entirely within the range to be expected and hardly can be used as a theoretical counterargument. The ionophoretic microtechnique has shown a number of other interesting results; it has confirmed the extremely critical localization of the chemoreceptors at and within

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

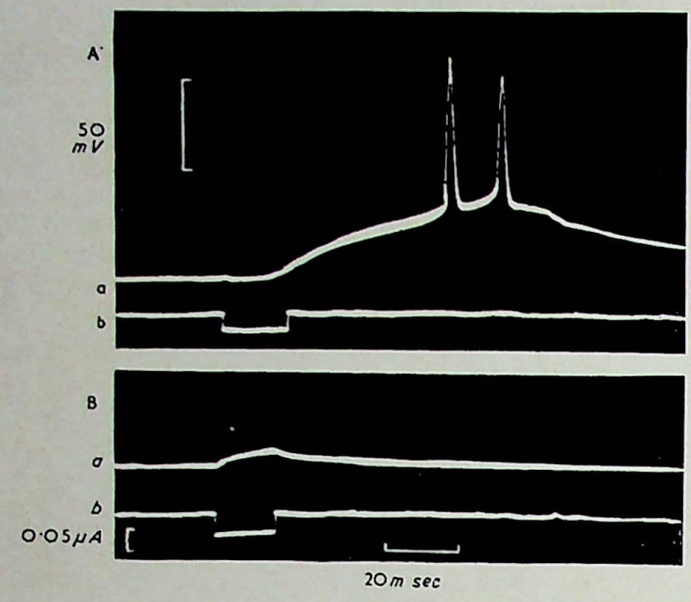


FIG. 1. External and intracellular application of acetylcholine to a motor end plate [from J. del Castillo and B. Katz, J. Physiol. 128, 157 (1955)]. Intracellular recording of membrane-potential changes from the junctional region of a frog muscle fiber. In *A*, an ACh-filled micropipette was placed on the outside of the end plate, and a quantity of ACh was released by passing a brief outward-directed current pulse through the pipette (registered in *b*). It produced the effect shown in trace *a*: a depolarization developing after a diffusion delay and culminating in two spikes. Between records *A* and *B*, the ACh-pipette entered the muscle fiber. An outward pulse produces now a small, direct potential change, owing to the passage of current through the fiber membrane. (If, for comparison, a KCl-filled micropipette is used, no potential change is produced by the pulse until the pipette has entered the fiber, when the effect is identical with that recorded in *B*, a.)

the neuromuscular junction. Moving the tip of the pipette by several microns can substantially reduce the effect of a given dose. Furthermore, it has been possible (Fig. 1) to insert the tip into the interior of the muscle cell, and so apply the acetylcholine alternatively to the external and internal side of the postsynaptic end-plate membrane.[4] The result showed that a depolarization was produced only by external, not by intracellular application; and this was observed with acetylcholine as well as carbamylcholine (a substance of similar action, but not destroyed by the local cholinesterase), and the same result was found for the blocking action of curarine. It seems that the first chemical attachment on to the receptor molecules must take place at the external surface of the end-plate membrane, which is, of course, the side facing the nerve endings from which the acetylcholine emerges under normal conditions.

Before discussing other peculiarities of the neuromuscular junction, the principal features and problems inherent in chemical transmission in general may be considered briefly.

There are two main steps interposed between the arrival of a presynaptic and the departure of a postsynaptic impulse: (i) the process by which the arriving impulse releases the transmitter substance, from its storage place inside the terminal into the narrow cleft between the contacting cells—this is a special case of what has been called "neuro-secretion"; (ii) there is the process by which the transmitter substance becomes attached to specific molecules in the postsynaptic cell surface and causes its electric membrane properties to change—this is a special example of *chemoreceptor* action, that is a process analogous to that occurring in our various chemical sense organs where a minute concentration change of some specific substance is registered in the form of sensory nerve impulses. As an intermediate step, one should consider also the mechanism by which transmitter molecules are transported across the small synaptic gap; however, the path length is only a fraction of a micron, and the time taken up by simple diffusion over such a short distance is well within the range of the observed synaptic latency.

EXCITATORY AND INHIBITORY CHEMORECEPTOR ACTION

To begin with, consider the second step, that is, the chemoreceptive mechanism. How do transmitter substances alter the membrane potential? Only a very incomplete answer to this question can be given. In general, the primary action leads to the opening of some ionic permeability channel in the membrane. Depending upon the size or specificity of this ionic channel, the membrane potential either tends to *fall* toward a low level, well beyond the firing threshold of an impulse, or it may become stabilized in the vicinity of the resting level or even tend to rise somewhat (hyperpolarize). In the first case, excitation ensues; in the last cases, one obtains an opposite, inhibitory, action (see Eccles[5]). But it may be noted that, underlying all of these changes, there is a common primary effect—namely, an increase of some ionic conductance.

For example, at the motor end plate there is evidence that acetylcholine causes the membrane permeability to increase, simultaneously, to several monovalent cations (e.g., sodium, potassium, ammonium) and possibly opens up an indiscriminate aqueous channel to all small ions on either side of the membrane. The result is a depolarization which has a "null point" at about 10 to 20 mv, negative inside, which corresponds to the level of a free-diffusion or liquid-junction potential between cytoplasm and external fluid. The effect is to short-circuit and depolarize the surrounding muscle membrane beyond the level at which a new impulse arises which then travels rapidly along the whole length of the muscle fiber. The methods by which these conclusions were reached have been described elsewhere;[6,7] briefly, they consisted in measuring the current/voltage relation across the end-plate membrane, and observing the particular level of the membrane potential at which the electromotive effects of ACh reversed, with normal as well as with altered composition of the ionic environment.

While ACh has a depolarizing and excitatory effect at the end plate, it produces the opposite action, that is

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

hyperpolarization or stabilization of the resting potential, in the regions of the heart muscle onto which it is released by impulses in the vagus nerve. Here also, the basic effect is an increase of ionic conductance, but the channel which is being opened is restricted to potassium, and does not include sodium. As a result the membrane tends to move towards, or to be held at, the potassium equilibrium potential which is usually somewhat greater than the existing resting potential (hence, a hyperpolarization).

Somewhat similar changes appear to be associated with the excitatory and inhibitory synapses in the motor neurons of the spinal cord.[5] At these junctions, the transmitter substances are unknown, but the "null points" of the potential changes which they produce have been determined and correlated with the existing ionic concentration gradients. Here also, an inhibitory hyperpolarization occurs which appears to be associated with an increase of membrane conductance to various small ions but excluding sodium; while excitation apparently arises from an indiscriminate "short circuit" in which sodium as well as the other small ions are allowed to pass.

Although the transmitter effects are fairly well understood in terms of ionic conductance changes, and the subsequent steps leading to excitation or inhibition present no special problem, the molecular mechanism by which the chemical attachment, e.g., of acetylcholine, to the receptive sites of the membrane alters its permeability is still far from understood. At one time it was thought possible that ACh^+ ions might produce a local depolarization without permeability change, simply by being able to move very rapidly through specific channels into the interior of the muscle fiber. This idea had to be abandoned when it became clear that the transfer of Coulombs during the end-plate depolarization exceeds the charge on the available ACh ions by several orders of magnitude. The observation shown in Fig. 1—namely, that a positive quantity of charge applied directly into the interior of the cell is much less effective in depolarizing the fiber membrane than an equivalent charge of ACh ions applied on the outside—illustrates this point rather clearly: most of the externally released ACh ions will diffuse away and have no chance of penetrating, or even colliding with, the end-plate surface. Yet under its influence much more positive charge enters the fiber than with the direct intracellular discharge. The conclusion is that the relatively few ACh ions released from the nerve terminal cause a vastly greater quantity of other, ambient, ions to flow through the end-plate membrane, and so achieve the great amplification of local current which is needed to transmit an impulse from the minute nerve endings to the much larger muscle cell.

Regarding the molecular combination between ACh and receptors and the subsequent chain of events, all that can be said at present is that a study of various chemical inhibitors (e.g., del Castillo and Katz[8]) suggests the presence of a 2- or 3-stage process whose kinetics resemble those of many enzyme-substrate reactions. Substances like tubo-curarine appear to act as competitive inhibitors, by interfering with the initial site of attachment, without themselves leading to the next phase which involves a change in the physical membrane properties.

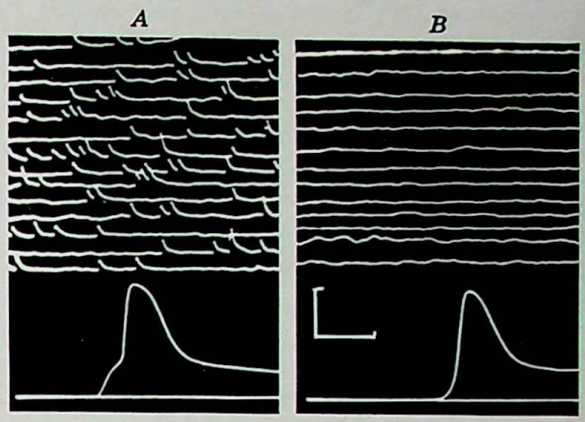


FIG. 2. Spontaneous miniature end-plate potentials. *A:* A recording microelectrode was placed inside a frog muscle fiber at the nerve-muscle junction. *B:* The electrode was placed 2 mm away into the same muscle fiber. The upper portions were recorded with slow speed and high amplification and show the occurrence of spontaneous small potential changes, restricted to the junctional region (calibrations: 3.6 mv and 47 msec). The lower portions show the response to a nerve impulse with fast-speed and low-gain recording (calibrations: 50 mv and 2 msec). The stimulus was applied to the nerve at the beginning of the trace; response *A* (at the end plate) shows the step-like initial end-plate potential which leads up to the propagating wave; response *B* shows only the propagated action potential, delayed by conduction over a distance of 2 mm [from P. Fatt and B. Katz, J. Physiol. 117, 109 (1952)].

Quantal Nature of Acetylcholine Release

An interesting feature which has emerged during a detailed study of the vertebrate nerve-muscle junction is that the release of ACh from the motor nerve terminals occurs in discrete packets or quanta each containing a large number of molecules. Even in the absence of a nerve impulse, such packets are released "spontaneously" at infrequent random intervals (Fig. 2). The arrival of an impulse at a cell junction apparently causes a few hundred events to be synchronized within a fraction of a millisecond instead of going on at a leisurely average rate of about 1/sec.[9]

The evidence for this state of affairs was obtained soon after it became possible to apply intracellular recording electrodes to the motor end plate. If a recording electrode is inserted into a resting muscle fiber well away from its junctional region, one observes a steady resting potential of about 90 mv, negative inside. But as one approaches the end plate with the recording probe, a characteristic form of spontaneous activity shows up which consists of an intermittent random

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

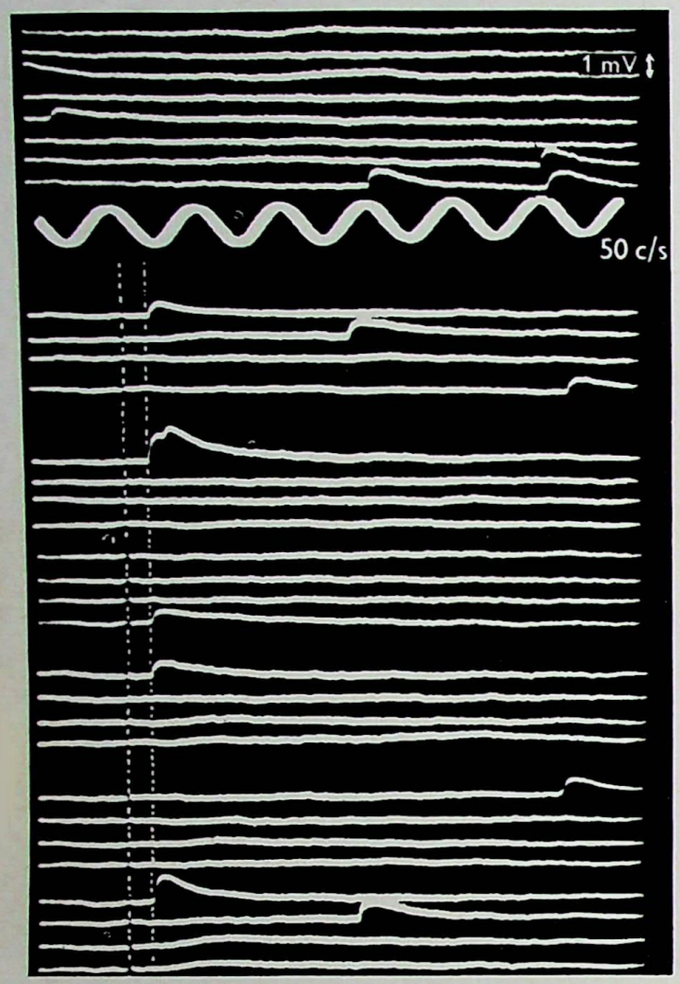


FIG. 3. Quantal units of end-plate response, recorded intracellularly from a muscle fiber in a calcium-deficient and magnesium-rich medium. The top portion shows a few spontaneous potentials. The lower part (below the 50-cps time signal) shows, in addition, the responses to single nerve impulses. Stimulus and response latency are indicated by a pair of dotted lines. There was a high proportion of failures, and only 5 single unit-responses to twenty-four impulses [from J. del Castillo and B. Katz, J. Physiol. 124, 560 (1954)].

discharge of minute potential changes of standard size and time course. Each deflection is a transient depolarization of the order of 0.5 mv, with a rapid (1 msec) rise and a slower decay, lasting altogether about 20 msec. It resembles in many respects the end-plate potential (e.p.p.)—that is, the immediate depolarization of the end plate produced after the arrival of a nerve impulse, but differs from it in its much smaller size (about 1%) and its spontaneous random occurrence. Fatt and I called it the miniature end-plate potential. It resembles the e.p.p. in its time course, its restricted localization to the innervated region of the muscle fiber, and its pharmacological reactions. It is reduced in size by curare, and its amplitude and duration increases when the local hydrolysis of ACh is prevented by a potent anti-esterase. In both respects, the miniature potential behaves exactly like the depolarizations produced by an applied dose of ACh, and we believe, therefore, that the spontaneous discharges arise from local random impacts of ACh on to the end-plate receptors. The source of these impacts is evidently a spontaneous release or leakage of ACh from the motor-nerve terminal where the substance is stored, for the miniature potentials vanish in the course of experimental nerve degeneration, at a time when neuromuscular transmission fails.

We considered the possibility that random molecular diffusion of ACh from the motor-nerve ending might be responsible for the minature e.p.p.'s. If this were true, then the same type of discharge ought to be elicited, at vastly increased frequency, by applying ACh to the fluid surrounding the end plate. This, however, is not the case: the depolarization which one then observes is continuously graded in size and time course, depending upon the dose and length of the diffusion path. It is clear that the effects of single molecular impacts of ACh must be far below the resolving power of our re-

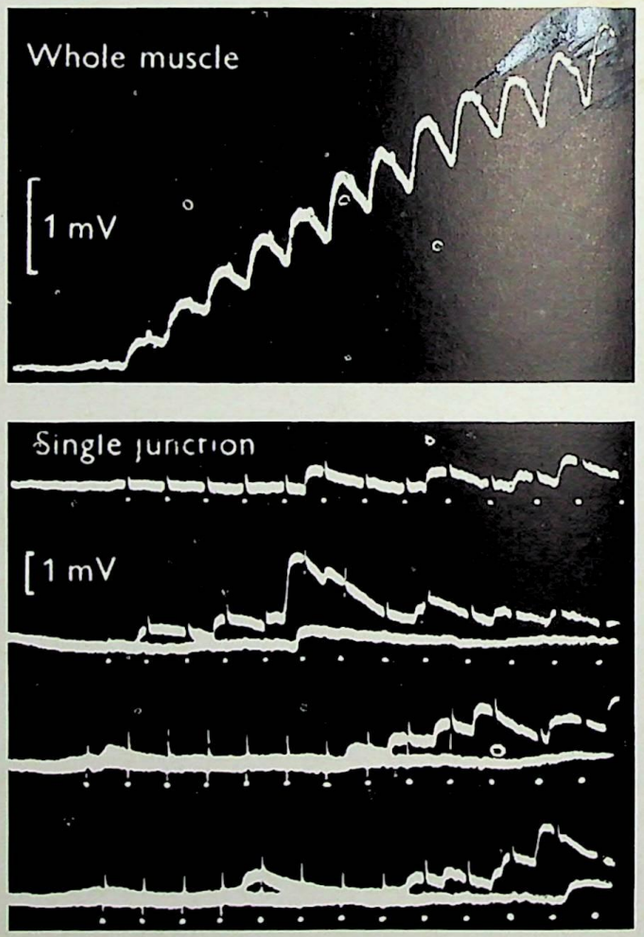


FIG. 4. Statistical properties of the end-plate response. The nerve-muscle preparations were blocked by a high magnesium- and calcium-deficient medium. The nerve was stimulated at 100 shocks per sec which produces a progressively increasing end-plate response. The upper record was obtained from the surface of a sartorius muscle showing the "smooth" average population response of a few hundred end plates. In the lower part, the response of a single end plate is recorded intracellularly, showing the quantal fluctuations of the response. Stimuli indicated by dots. Note spontaneous potentials on the superimposed "base lines" [from J. del Castillo and B. Katz, J. Physiol. 124, 574 (1954)].

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

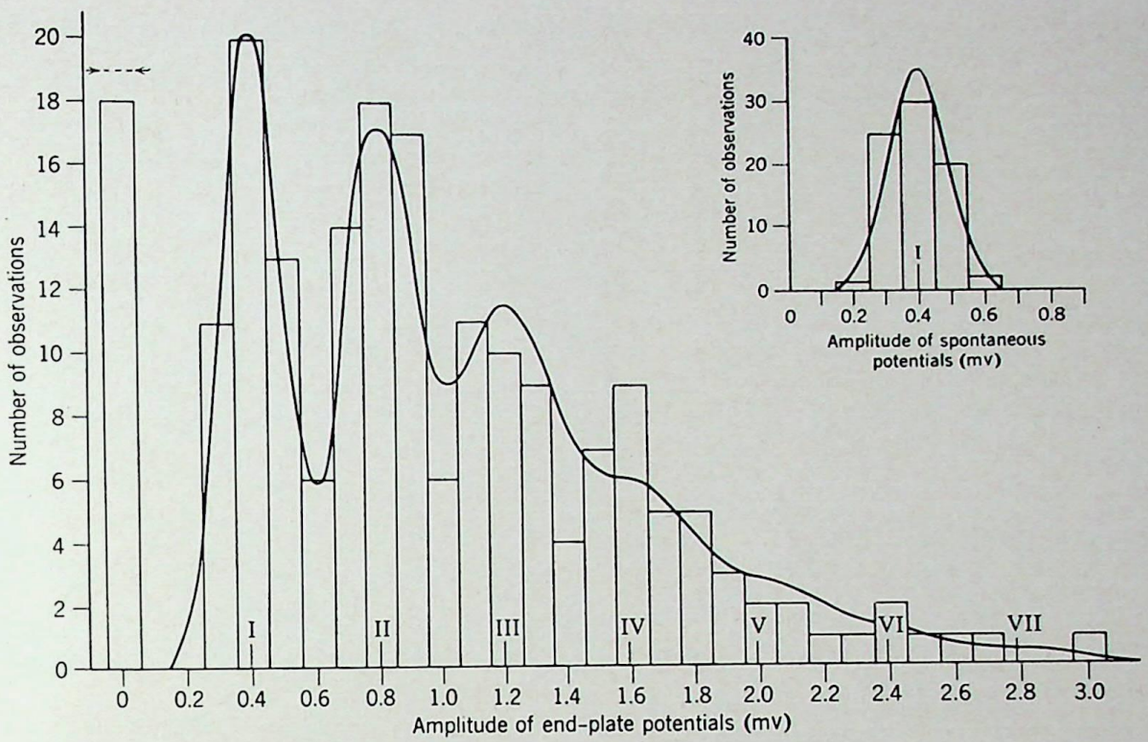


FIG. 5. Histograms of e.p.p. and spontaneous potential amplitudes (inset), from a mammalian end plate blocked by magnesium. Peaks of e.p.p. amplitude distribution occur at 1, 2, 3, and 4 times the mean amplitude of the spontaneous miniature potentials. A Gaussian curve has been fitted to the latter and used to calculate the theoretical distribution of e.p.p. amplitudes (continuous curve). Arrows indicate expected number of failures (zero amplitude) [from I. A. Boyd and A. R. Martin, J. Physiol. **132**, 74 (1956)].

cording equipment, and conversely that a discrete miniature e.p.p., with its standard size and brief time course, must be due to a synchronous package of ACh molecules, may be hundreds or thousands, discharged in the immediate vicinity of our end-plate receptors.

That this packet is the basic coin in which the trans-

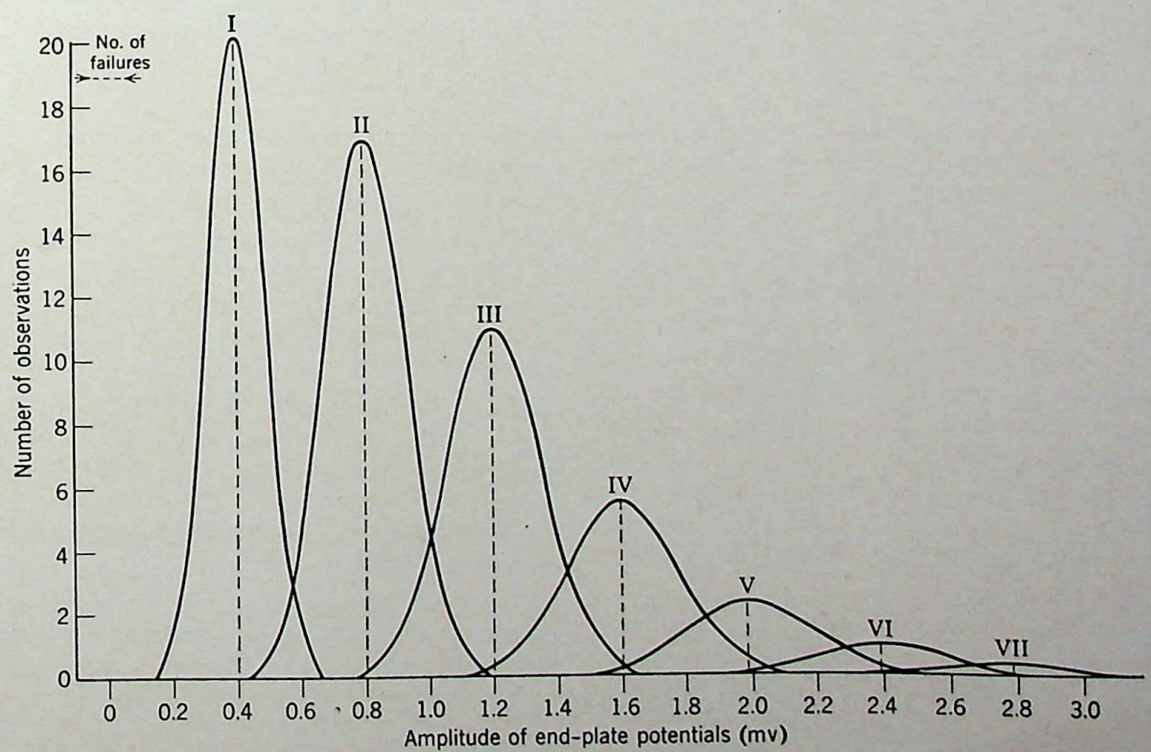


FIG. 6. Method of obtaining the continuous theoretical curve in Fig. 5. A Poisson distribution was calculated, for a mean value m = mean amplitude of e.p.p. responses/mean amplitude of spontaneous potentials. The calculated numbers of each Poisson class have been distributed along Gaussian curves, corresponding to multiples of the spontaneous potentials (see Fig. 5). Algebraic summation of ordinates gives the continuous curve of Fig. 5 [from I. A. Boyd and A. R. Martin, J. Physiol. **132**, 74 (1956)].

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

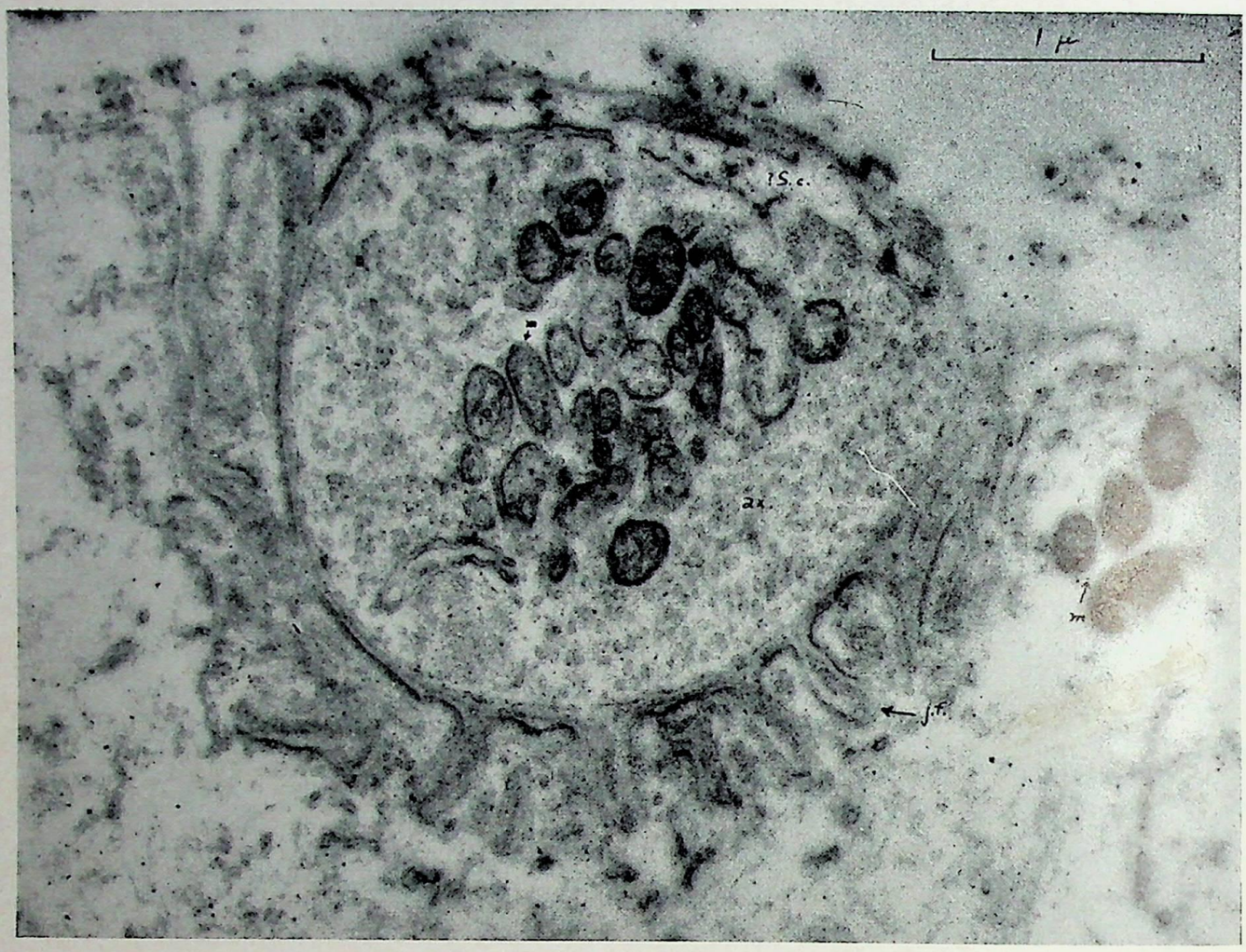


FIG. 7. Electron micrograph of a reptilian neuromuscular junction. Diameter of the motor-nerve ending is approx 2.5 μ; it contains, in addition to mitochondria (large dark particles), many small vesicles of a few hundred A units diameter [from J. D. Robertson, J. Biophys. Biochem. Cytol. 2, 381 (1956)].

mitter is normally delivered from the nerve endings, has been shown in several kinds of experiments. The release of ACh by an impulse depends upon a number of "co-factors" which can be varied experimentally: among these perhaps the most important are the concentrations of calcium and magnesium in the surrounding fluid. Calcium is an essential adjuvant, magnesium an inhibitor of the release mechanism. By lowering the calcium and raising the magnesium concentrations, the quantity of ACh liberated during an impulse, and the size of the resulting e.p.p., can be progressively reduced towards zero. The point of interest is that, during such an experiment, the e.p.p. at each individual end plate is found to be diminished in discrete steps, which correspond to the dropping out of individual miniature potentials, one by one. With a suitable ratio of Ca/Mg concentrations the liberation can be reduced to a small number of ACh packets;[10] in this condition, the size of the end-plate potential during successive nerve impulses has been found to fluctuate in a characteristic stepwise manner, corresponding to a Poisson-wise distribution of the number of units released in each instance.[11,12] The unit-step is identical with the spontaneously occurring miniature potential. Examples are shown in Figs. 3–5, Fig. 6 showing the method whereby Fig. 5 is obtained; detailed studies of this effect on a variety of vertebrate nerve-muscle junctions have made it certain that transmission is brought about by a summation of many of these quantal units of activity.

A further point of interest is that the size of the unit parcel of ACh which is delivered from the nerve endings, spontaneously or in response to an impulse, is relatively constant at all cell junctions and apparently quite unaffected by the many changes which may be imposed on the system in the course of an experiment. On the other hand, the probability of release of any one parcel in a given time interval—that is, the frequency of the miniature potentials—can be altered by several orders of magnitude, for example, by electrically depolarizing the nerve ending with a steady current, or by changing the chemical composition of the environment. The action of the nerve impulse itself can be described as causing a momentary enormous increase in the frequency of the miniature potentials (by a factor of nearly 10^6, provided a high Ca/Mg ratio exists in the surrounding medium).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

The basis of this multimolecular quantum of ACh release is not yet known; an attractive suggestion is that the transmitter substance is stored, within the nerve endings, in minute intracellular corpuscles from which it is discharged at the surface in an all-or-none manner.[6,7] Electron micrographs[13,14] have revealed a mass of fairly densely packed so-called vesicles inside the nerve terminals, and it is conceivable that these are the intracellular bags in which ACh is being stored prior to its release (Fig. 7). It is possible to imagine mechanisms by which a collision between such particles and certain critical spots in the nerve membrane could bring about sudden liberation of the vesicular contents straight into the synaptic cleft. But while it is easy to present speculations which are compatible with the existing evidence, to put them on a firm experimental basis will be a much more difficult task.

BIBLIOGRAPHY

[1] E. J. Furshpan and D. D. Potter, Nature **180**, 342 (1957).

[2] T. H. Bullock and S. Hagiwara, J. Gen. Physiol. **40**, 565 (1957).

[3] S. Hagiwara and I. Tasaki, J. Physiol. **143**, 114 (1958).

[4] J. del Castillo and B. Katz, J. Physiol. **128**, 157 (1955).

[5] J. C. Eccles, *The Physiology of Nerve Cells* (The Johns Hopkins Press, Baltimore, Maryland, 1957).

[6] J. del Castillo and B. Katz, Progr. in Biophys. and Biophys. Chem. **6**, 121 (1956).

[7] B. Katz, Bull. John Hopkins Hosp. **102**, 275 (1958).

[8] J. del Castillo and B. Katz, Proc. Roy. Soc. (London) **B146**, 369 (1957).

[9] P. Fatt and B. Katz, J. Physiol. **117**, 109 (1952).

[10] J. del Castillo and B. Katz, J. Physiol. **124**, 560 (1954).

[11] J. del Castillo and B. Katz, J. Physiol. **124**, 574 (1954).

[12] I. A. Boyd and A. R. Martin, J. Physiol. **132**, 74 (1956).

[13] J. D. Robertson, J. Biophys. Biochem. Cytol. **2**, 381 (1956).

[14] G. E. Palade, Anat. Record **118**, 335 (1954).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

57

Some Quantifiable Aspects of the Electrical Activity of the Nervous System (with Emphasis upon Responses to Sensory Stimuli)*

WALTER A. ROSENBLITH

Department of Electrical Engineering and Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge 39, Massachusetts

IN assessing the sensory performance of higher organisms (Rosenblith, p. 485), a certain number of characteristics emerge that suggest problems for a quantitative study of the electrical activity of the nervous system. In considering organisms engaged in communication tasks, it has been seen (a) that their performance is statistical in character; (b) that they need more time to handle more information; (c) that their capacity to discriminate and their speed of reaction depend on stimulus intensity; and (d) that their repertory of absolute identifications ("span of absolute judgment") is relatively small and seemingly based upon the ability to make several rather crude discriminations simultaneously, i.e., to classify environmental sensory inflow into some rather gross categories. One is thus led to examine the electrical phenomena in the nervous system in their statistical aspects and to look for over-all, i.e., relatively gross patterns of electrical behavior such as those exhibited by populations of neurons at various levels of a sensory system, rather than to attach disproportionate significance to the behavior of isolated members of such populations. By choosing this quasi-thermodynamic approach, it is not intended to deprecate the complementary view that derives its inspiration from statistical mechanics. But, it must be emphasized that we do not know how to sample a population of neural units adequately (it is far from easy to define a "typical" unit when experimenting with a single microelectrode), and that there is much to be learned of the laws of interaction between units that presumably belong to a population.

It is from this viewpoint that data are presented in this article on (a) electrical responses evoked by sensory stimuli and on (b) so-called on-going activity; the latter may be thought of as reflecting an organism's "state" as it reacts selectively to sensory stimulation.

SCHEMATIC FLOW CHART OF A MODEL SENSORY SYSTEM

It is obviously impossible to deal here realistically with the neuroanatomy of those structures in the nervous system that become involved when a sensory stimulus is presented. It may, therefore, be expedient to start the discussion by presenting, in Fig. 1, a highly oversimplified view of a model sensory system. *Pour fixer les idées*, consider the auditory system, which is perhaps the one for which the greatest amount of quantitative data is available.

A stimulus impinges upon the organism: The appropriate transducer (here located in the inner ear) is activated, and electrical activity is generated in the region of the hair-cell-neural junction.[3] If the stimulus is a "transient," certain subsequent electrical phenomena at the various higher levels of the nervous system are comparatively easy to follow, and both amplitude and temporal characteristics of responses at various locations can be measured with accuracy.

Once the electric signal has entered the organism's nervous system, it is possible for it to ascend the classical afferent pathway to the cortex. The route can be represented schematically by straight segments in the nature of nerve tracts; these alternate with circles that symbolize the main relay stations of the pathway. Along the axons of the longitudinal tracts, the transmission of signals is predominantly in the form of conducted nerve impulses (all-or-none activity or "spikes"). Transmission is one-to-one in the sense that a given axon exhibits identical patterns of "spikes" along its entire length. The relay stations exhibit activity that has both "discrete" and "continuously graded" aspects. It is in these regions that both excitatory and inhibitory synaptic junctions play an important integrative role by permitting interactions of the many-to-one and perhaps also one-to-many type.

One of the more significant features of the flow chart of Fig. 1 is that it depicts, in parallel with the classical afferent pathway, another route that leads—via the brain-stem reticular formation—to the cortex. In his recent book,[4] Magoun characterizes the reticular system as follows:

"From recent study of the brain stem . . . the concept has developed of a major non-specific or transactional mechanism in the brain, paralleling the specific sensory and motor systems of classical neurology and richly interconnected with them. This non-specific neu-

* This work was supported in part by the U. S. Army (Signal Corps), the U. S. Air Force (Office of Scientific Research, Air Research and Development Command), and the U. S. Navy (Office of Naval Research).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

ral system is distributed through much of the central core of the brain stem and, as spokes radiate from the hub of a wheel to its peripheral working rim, so functional influences of this central system can be exerted in a number of directions: caudally upon spinal levels to influence both reflex and other spinal activity; rostrally and ventrally upon hypothalamic and pituitary mechanisms through which influences can be exerted upon visceral and endocrine functions; cephalically upon the diencephalic and rhinencephalic brain, where affect and emotion now reign instead of in the heart; and, more cephalically and dorsally still, upon the neocortex of the cerebral hemispheres which, with its interconnected thalamic and basal ganglionic masses, subserves higher sensory motor and intellectual performance. The influences of this non-specific system in the brain stem are thus brought to bear upon most other portions and functions of the central nervous system, either to diminish or to raise the level of their activity, or to interrelate or integrate their several performances." . . . "It does this as a reflexion of its own internal excitability, in turn a consequence of both afferent and corticifugal neural influences, as well as of the titer of circulating humors and hormones which affect and modify reticular activity."

Relatively few of these properties are depicted in Fig. 1. Yet there can be seen collaterals that branch off from the afferent pathway to furnish the reticular system with information regarding the signals that are being processed along the more direct and specific route. There is also an indication that this system receives impulses from all sensory modalities, i.e., both from the *milieu extérieur* and the *milieu intérieur*. And there are indications that the flow of neural signals for even a single sensory modality is not just one-way in the reticular system, i.e., directed toward the cortex; there are important descending pathways whose functional significance is just beginning to be understood.

It is perhaps somewhat surprising to learn that neuroanatomists have given descriptions of the reticular system before the beginning of the twentieth century, but that it was only after Bremer's experiments of the midthirties, and in particular after the now classical Moruzzi-Magoun experiments[5] on the "Brain stem reticular formation and activation of the EEG" (1949), that this system began suddenly to loom large in the understanding of all integrated behavior. Previous to that period (and even in a large number of experiments performed since then, in the fifties), the use of centrally acting anesthetics has resulted in differential interference with reticular activity. Under these circumstances, most explanations of the role which the nervous system plays in perception or in the handling of sensory information are based on evidence from, neurophysiologically and behaviorally speaking, badly distorted preparations.

Among the more dramatic findings that have forced one to give up oversimplified views of the organization

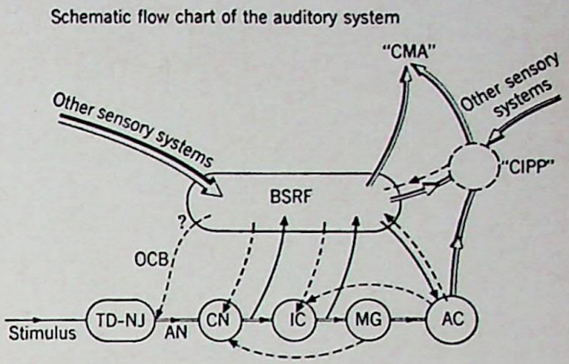


FIG. 1. For the sake of simplicity, this diagram presents only a gross outline of the neuroanatomy of the auditory system. Such important way stations as the cerebellum, the superior olive, the lateral lemniscus, and other regions of the brain from which responses to acoustic stimuli have been recorded have been left out; the still incompletely known and complex organization of the brain-stem reticular formation (BSRF) has not even been suggested. Furthermore, no attempt has been made to indicate connections that correspond to acoustic reflexes of short latency: middle ear muscles contract about 10 msec after stimulus delivery, while eye blinks occur about 35 to 40 msec after delivery of a strong click; such reflexes represent simple forms of integrated behavior that do not necessarily involve the higher levels of the auditory system.

In order to appreciate the number of neural elements present along the afferent pathway, Chow's[1] numerical estimates of the auditory system of rhesus monkey are reproduced here. There are about 10^5 cells in the cochlear nucleus (CN), about 4×10^5 cells in the inferior colliculus (IC) and also in the medial geniculate (MG); the auditory cortex (AC) is estimated to contain about 10^7 cells; these numbers, which represent orders of magnitudes only, may be compared with the 3×10^4 fibers in the auditory nerve (AN) and the 2.5×10^4 hair cells in the inner ear, i.e., at the transducer neural junction (TD-NJ). The foregoing data are the results of histological investigations; hence, they do not permit any direct inference concerning the number of neural units that will be either directly excited or inhibited in the presence of a stimulus of given intensity.

The remaining symbols are to be identified as follows: OCB stands for the olivo-cochlear bundle (depicted in dashed lines like the other descending pathways) whose influence upon the activity of the auditory nerve has been investigated by Galambos,[2] CIPP stands for the central information-processing pool of neurones, whose existence, though neither neuroanatomically nor neurophysiologically established, is conceptually useful. The relations of this hypothetical construct to the sensory-projection areas and association areas remain to be worked out to give a more realistic account of the role that the brain and its various components play in the organism's handling of sensory information. CMA stands for the inferred control unit of the motor activity that takes place in connection with the stimulus-related response.

and functioning of sensory systems have been the following: (a) sectioning of the classical auditory ascending pathway was shown to leave intact an animal's ability to be aroused by sound; (b) animals whose auditory cortex had been ablated were shown to be capable of simple frequency and intensity discriminations, though they were not capable of distinguishing more-complex patterns such as simple melodies. Experiments of this sort should, however, not be interpreted to mean that the anatomical structures that can be removed or interfered with play no role in the normal performance of sensory tasks. There are certainly different ways of assessing impairment of performance, and the mere statement that an animal is still capable of carrying out a given task, after ablation, does not force the conclu-

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

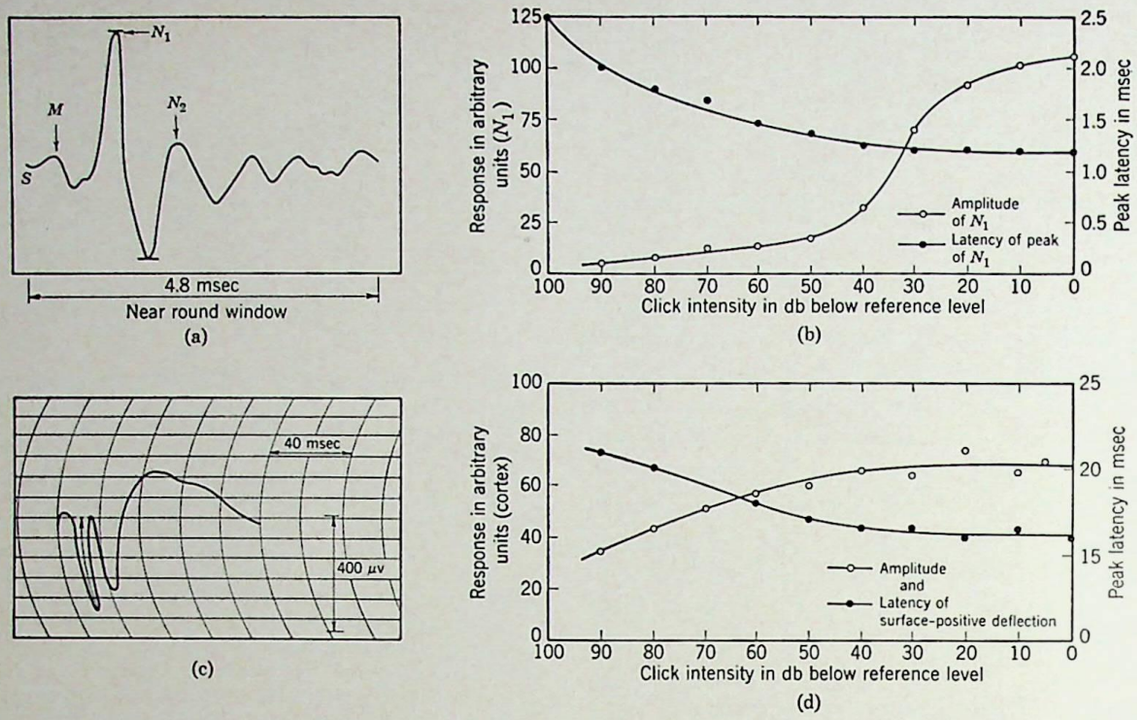


FIG. 2. (a) A typical response to a click of medium intensity recorded near the cat's round window of the inner ear. M is a so-called microphonic potential which is non-neural in origin. N_1 is the earliest and most prominent neural component that can be recorded at this location. It represents the summated action potentials from the auditory nerve. S indicates the occurence of the electrical pulse that generates the acoustic click. (b) An example of how the amplitude and latency of N_1 vary as the intensity of the click is varied. The latency values include a delay of approximately 0.5 msec that corresponds to the time that it takes for the sound to travel from the earphone diaphragm to the eardrum. (c) An average response to a click recorded from a location on the auditory cortex of an unanesthetized cat. The averaging operation was carried out by an electronic computing device.[14] Surface positive deflections are plotted downward. (d) Intensity functions for the baseline-to-peak amplitude and the peak latency of the surface positive component of the cortical response to clicks in an anesthetized cat.

sion that there is no sensory deficit. More generally, one is impressed with the fact that the nervous system of higher animals seems to be characterized by enough flexibility (which some people have metaphorically referred to as a "safety factor" or "redundancy") so as to permit the performance of certain rather basic discriminations in a variety of ways.

For a long time students of the nervous system have wondered how its flexibility manages to deal with sensory stimuli, when they are in the focus of the organism's attention in contrast to situations in which they seem to be of no importance. Adrian formulated this question most incisively at the 1954 Symposium on *Brain Mechanisms and Consciousness*[6]:

"The operations of the brain seem to be related to particular fields of sensory information which vary from moment to moment with the shifts of our attention. The signals from the sense organs must be treated differently when we attend to them and when we do not, and if we could decide how and where the divergence arises, we should be nearer to understanding how the level of consicousness is reached. The question is whether the afferent messages that evoke sensations are allowed at all times to reach the cerebral cortex or are sometimes blocked at a lower level. Clearly we can reduce the inflow from the sense organs as we do by closing the eyes and relaxing the muscles when we wish to sleep and it is quite probable that the sensitivity of some of the sense organs can be directly influenced by the central nervous system. But even in deep sleep or coma there is no reason to believe that sensory messages no longer reach the central nervous system. At some state therefore on their passage to consciousness the messages meet with barriers that are sometimes open and sometimes closed. Where are these barriers, in the cortex, the brainstem, or elsewhere?"

The schematic flow diagram of Fig. 1 suggests, in agreement with the beautiful experiments that have been carried out since 1954, that the barriers that Adrian refers to are at no single place in the nervous system, but that they are multiple. This is clearly not the place to review in detail the experiments carried out by Hernández-Peón[7] and many others.[8–13] Suffice it to report that all of these experiments indicate that the descending pathways exercise powerful gating, screening, and regulating influences. This central or centrifugal control of the afferent sensory inflow has been demonstrated to be effective across sense modalities (for example, the presence of a mouse or of fish odors reduces the response to a click at the level of the cat's cochlear nucleus) or during "habituation" (which presents itself as a reduction in the amplitude of evoked responses for stimuli that have been repeated over and over again). This effect of habituation can, however, be suddenly abolished by pairing the "boring" stimulus judiciously with a few electric shocks. From the available evidence

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

(these dramatic effects of attention and habituation are, for instance, not found in anesthetized animals), one is led to conclude that the selective processing of the sensory inflow is under the control of the brain stem or of descending pathways that are structurally or functionally closely related to the reticular system. Much remains to be learned about the precise ways in which the informational aspects of a sensory signal affect its progress in the nervous system, but it is already clear that sensory systems cannot be characterized by invariant sensitivity and invariant "tuning."

EVOKED RESPONSES ALONG THE AFFERENT PATHWAY

Having discussed the schematic flow chart and some of the phenomena that led to its formulation, it is now necessary to admit that most of the quantitative data that exist deal with electrical events along the classical afferent pathway. This admission should give one a more realistic appreciation of what the phenomena are for which there exist functional relations of the stimulus-response type. What follows should, therefore, be merely considered as an introduction to some problems of recording, data-analysis, and model-making in relation to the electrical activity found at various levels of the nervous system.

The left side of Fig. 2 depicts samples of responses to clicks from populations of neural units at the level of the auditory nerve and at the level of the auditory cortex. As indicated, it is possible to make quantitative measurements along both the voltage and the time axes of these displays. Thus, one can measure amplitudes of deflections (either peak-to-peak or baseline-to-peak) and "latencies" (i.e., time intervals that have elapsed since the delivery of the stimulus) of extreme values or zero crossings. Naturally, the characteristics of the recording equipment must be chosen appropriately. It would be futile if one were to attempt to measure accurately either the latency or the amplitude of rather sharp deflections on the ink tracings taken by the customary EEG recording device.

When such amplitudes or latencies are measured, one must ask whether one shall be satisfied with average values or whether one shall need to know entire distributions in order to be able to give a rational account for the neuronal events. As has been shown by McGill,[15] Macy,[16] Frishkopf[17] and others, evoked responses are afflicted with variability, especially in regard to amplitude and particularly at the higher neural centers unless the animal is rather heavily anesthetized. It is important to realize that this variability of evoked responses to identical stimuli is present in spite of the fact that the gross electrodes that record the activity of reasonably large neural populations are already performing an averaging process. Some studies of the variability of evoked responses in anesthetized preparations have been undertaken[16–18] and have yielded results that

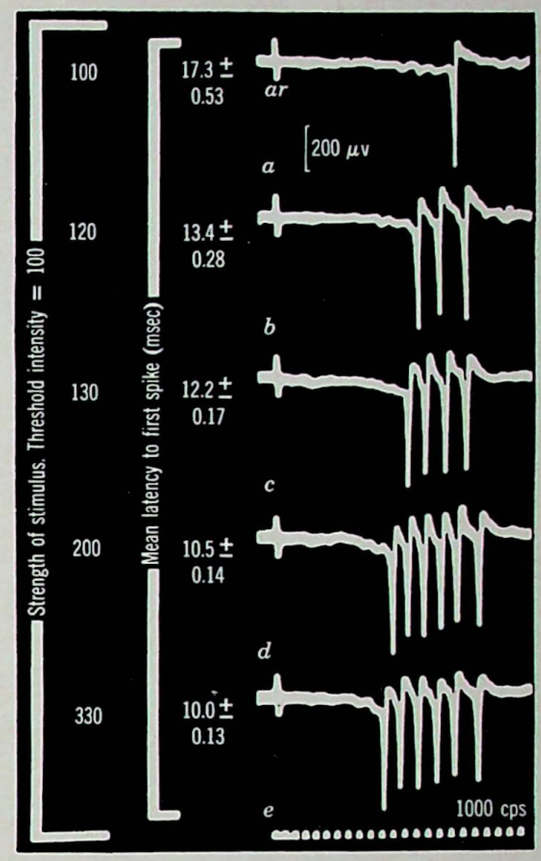


FIG. 3. Responses of single units in tactile thalamic region to transient electrical stimulation of forepaw. Position of stimulation is unchanged for all records. Numbers in column at extreme left give a measure of stimulus strength. Numbers in second column indicate (in msec) mean latency to first spike and standard error of the mean for each stimulus intensity. The records in the right-hand column depict the modal value of the number of spikes that are recorded for each intensity. The short vertical line at the left of each response trace indicates the delivery of the electrical stimulus [from J. E. Rose and V. B. Mountcastle, Bull. Johns Hopkins Hosp. 94, 238 (1954)].

are suggestive from the viewpoint of probabilistic models.

On the basis of amplitude and latency measurements, one can plot intensity functions, i.e., functions that relate the average amplitude of characteristic deflections and their average latency to the intensity of the stimulus. Figures 2(b) and 2(d) illustrate such functions. These and similar graphs permit one to state the following generalization: As the intensity of a transient sensory stimulus is increased, the average amplitude of evoked responses increases (or at least does not decrease), and the average latency of either the onset or the peak of the deflection decreases. [Parallel generalizations can, of course, be formulated for single neural units (see Figs. 3 and 4, from Rose and Mountcastle[19]).] The detailed shape of these intensity functions varies for different locations in the nervous system and for different stimuli, and it is, therefore, hardly appropriate to search for a unique representation of stimulus intensity in the nervous system.

It must be noted that the quantification process just described runs into difficulties if stimuli of appreciable

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

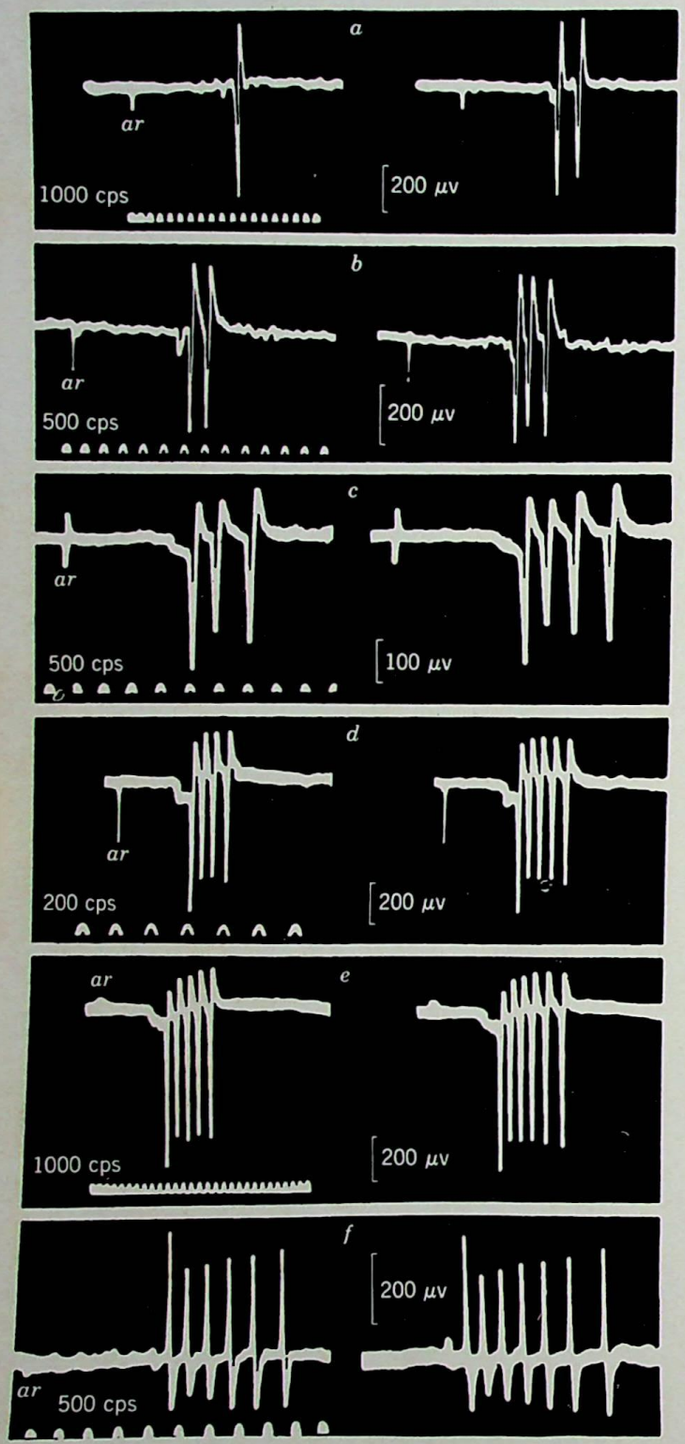


FIG. 4. Each pair of records comes from the same unit. All six units were activated by electrical stimulation of different locations on the skin. The stimulus artefect (*ar*) indicates the instant of stimulus delivery [from J. E. Rose and V. B. Mountcastle, Bull. Johns Hopkins Hosp. 94, 238 (1954)].

duration are used. Thus far, it has not been possible to suggest, especially for the higher neural centers, meaningful measures of the evoked activity on the basis of visual inspection only. It is by no means certain that simple intensity functions will be obtained for all instants during which the stimuli are "on"; the interaction of excitatory and inhibitory phenomena, as well as various adaptation phenomena, can be counted upon to complicate the picture.

Before dealing with microelectrode data on the behavior of single units at various way stations of the nervous system, mention should be made of the series of mathematical models developed by McGill,[15] Frishkopf,[17] Goldstein,[20] and Macy.[16] These models, which are all probabilistic in character, relate certain aspects of gross-electrode responses (i.e., responses to which populations of neural units have contributed) to stimulus parameters. All of these models make certain, physiologically not unreasonable, assumptions regarding the behavior of single units in the auditory nerve or at the auditory cortex. From these postulated properties and from certain assumptions regarding the interaction between various members of a population of neural units, the behavior of population responses (i.e., evoked responses recorded by gross electrodes) can be predicted and compared with experimentally available data. The response parameters that have been investigated include average amplitudes, variability in amplitude, wave form of a population response, and temporal recovery characteristics. The stimuli were clicks or bursts of noise: these were presented either singly or in pairs (or even triplets), either in the absence or in the presence of background noise. While it is obviously premature to forecast the ultimate success of such model-making, it is safe to assert that the models have demonstrated their usefulness in summarizing sizable quantities of experimental data and in suggesting systematic experimentation in an area wherein workers seem only too often satisfied by demonstrating the existence of phenomena. These models force one, furthermore, to face an issue that may well be critical for the understanding of the organization of the nervous system: What is the relationship between the behavior of the components that can be studied (single cells), the behavior of subsystems (made up of more or less homogeneous populations of single units), and the behavior of an entire sensory system? For it is hardly reasonable to expect that the knowledge of the functioning of the brain will be revealed *in toto* in the absence of a working knowledge of principles of organization of functional units of the brain.

PATTERNS OF ELECTRICAL ACTIVITY IN SINGLE UNITS OF SENSORY SYSTEMS

Over twenty years ago, first Blair and Erlanger[21] and then Pecher[22] showed that, under carefully controlled conditions, a single nerve fiber will sometimes respond and sometimes not respond when the same stimulus is presented repeatedly. They furthermore showed that, at least to a first approximation, spike responses in fibers that belonged to the same nerve trunk were statistically independent. In 1950,[23,24] it was shown that the probability of response to a click of a single unit in the cochlear nucleus of the cat increases from "almost never" to "almost always" as the stimulus energy was multiplied by a factor of approximately 100 (or 20 db);

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

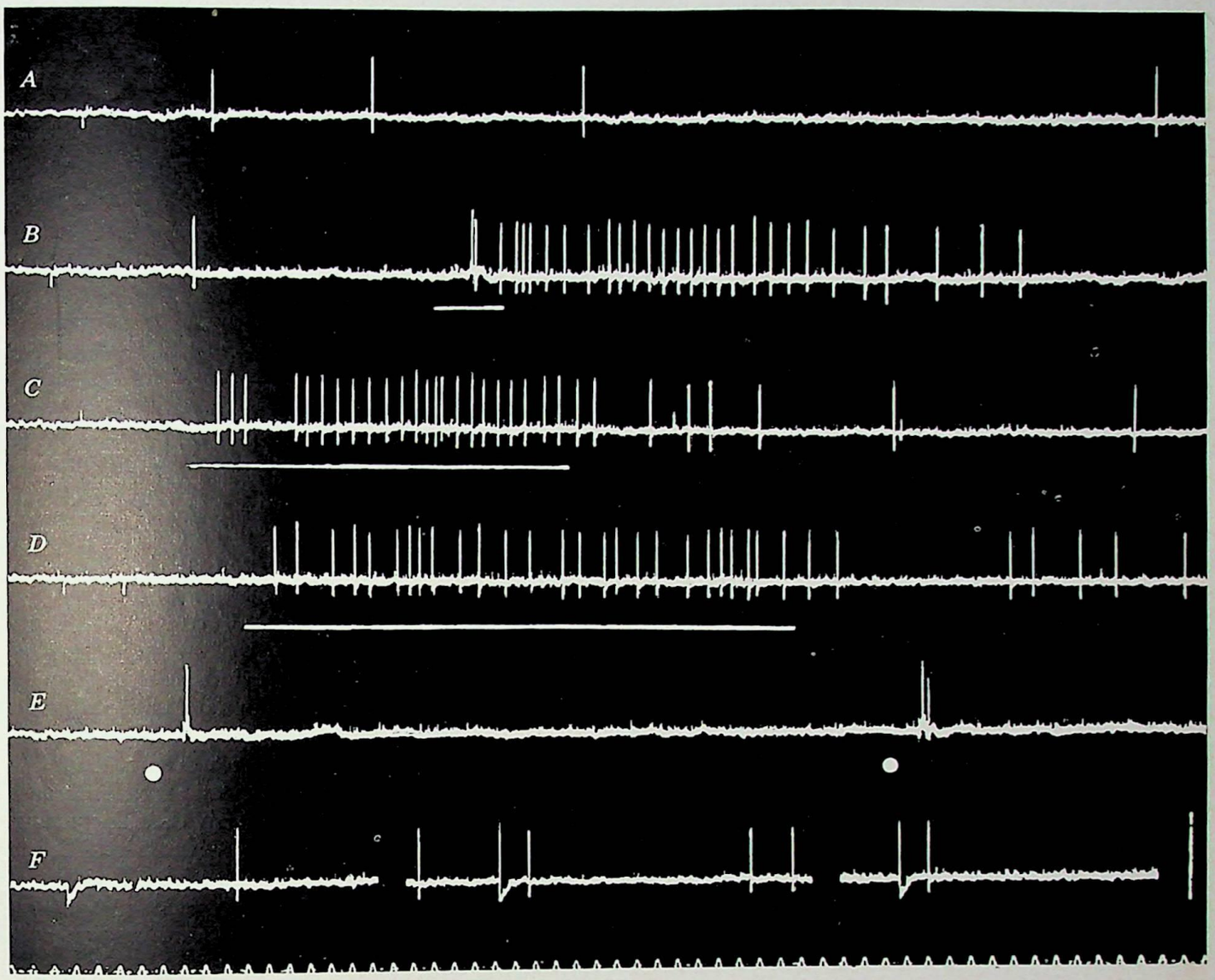


FIG. 5. Microelectrode recording from a unit in the medial part of the pontine reticular formation of a cat: *A*, spontaneous activity; *B*, tapping the ipsilateral forelimb; *C*, rubbing the animal's back; *D*, touching the cat's whiskers; *E*, hand claps; *F*, single shocks delivered to the animal's sensorimotor cortex. Delivery of stimuli is indicated by white horizontal lines (*B*,*C*,*D*), dots (*E*), and artifacts that displace the record's baseline (*F*). Time calibration at the bottom of figure indicates 10-msec intervals [from M. Palestini, G. F. Rossi, and A. Zanchetti, Arch. Ital. Biol. **95**, 97 (1957)].

at the same time, the latency of the spike response that is elicited decreases substantially. Since then, a considerable amount of such data has become available for different structures in the nervous system.[25,26]

Rose and Mountcastle[19] have given some of the most beautiful records that are available today. These records illustrate the finesse and the subtlety with which the nervous systems can "code" stimulus intensity into response patterns of single cells. These workers stimulated the skin of cats electrically and recorded the activity of single neurons in the thalamic relay nucleus of the cat's tactile system. Figure 3 shows how the number of spike responses increases and the latency of the first spike decreases as the strength of the transient peripheral stimulus is increased. Measurements of intervals between spikes yield such interesting results as: (a) Units start to fire at a rate which is higher, the larger the number of spikes in a train of repetitive discharges. (b) The interspike interval increases throughout a train and firing ceases when the interval exceeds a value such as 2 msec.

Figure 4 shows, in each horizontal subdivision, pairs of responses from the same unit to identical stimuli. These discharges are again all-or-none in character and, in spite of the fact that their height (amplitude) varies within a given train, there is overwhelming evidence that all spikes in a single train come from the same unit. This figure furnishes striking evidence for the view according to which the variability in response patterns of single units to identical stimuli renders their statistical description imperative.

Figures 3 and 4 represent locations in the modality-specific classical afferent pathway. The next record, taken in Moruzzi's laboratory (Fig. 5, from Palestini *et al.*[27]), illustrates, in contrast, the response pattern of a single unit in the brain stem.[28] This unit was responsive

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

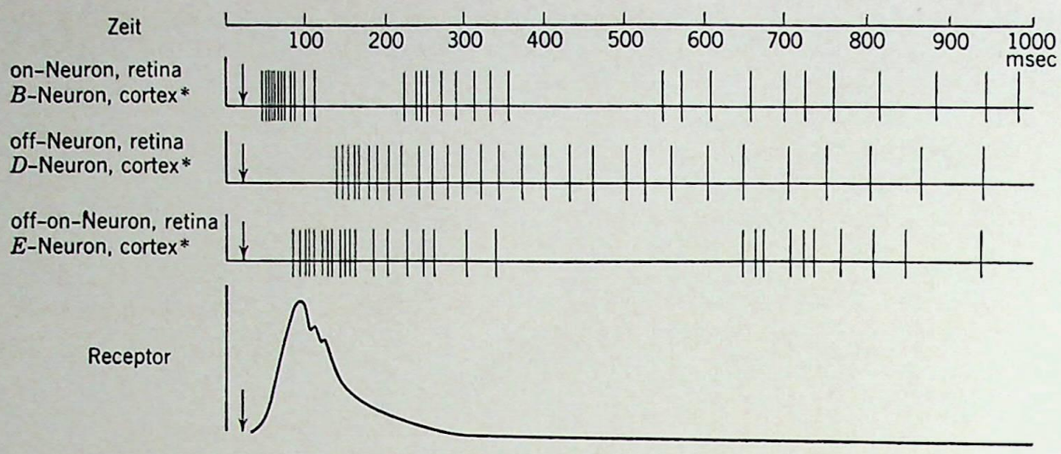


FIG. 6. Typical patterns of unitary discharges in the cat's visual cortex in response to a short flash of 3000 Lux; the bottom trace shows a typical electroretinogram that is observed at the same time. Arrow indicates delivery of flash; the *B*-, *D*-, and *E*-neurons reflect, to a certain degree, patterns of retinal excitation (on, off, and on-off). The pattern of activity of the *B*-neuron—i.e., an initial burst followed by a pause of about 0.1-sec duration—is found in other sensory systems (after O. J. Grüsser and A. Grützner[29]).

to all test stimuli that were tried, although the pattern of spike activity was rather different for the different modalities.

The next microelectrode record (Fig. 6) demonstrates that some of the response patterns that can be observed in sensory systems are discernible over time intervals whose length is comparable to behavioral reaction times. Jung and his group[30] distinguish four different types of neurons in the visual cortex on the basis of their activity patterns. Figure 6 describes schematically three of these typical patterns. The fourth pattern, which corresponds to the so-called *A*-neuron, is not represented. Actually, more than one-third of the units in the visual cortex from which records have been obtained fail to yield visually detectable changes in their spontaneous activity after delivery of an optical stimulus.† There is, however, some evidence that these units may exhibit some kind of responsiveness, provided a form of optical stimulation is used that attracts the animal's attention..

From the highly selected evidence that has been presented in Figs. 3 to 6, a certain number of conclusions impose themselves regarding the strategy of future model-oriented research in this area. New recording devices and new techniques of analysis are needed. One must be able to record simultaneously the firing patterns of not one or two cells but of significant assemblies and subassemblies of cells. Gross electrodes permit the recording of potentials that are very coarse spatio-temporal averages of both spike and continuously graded

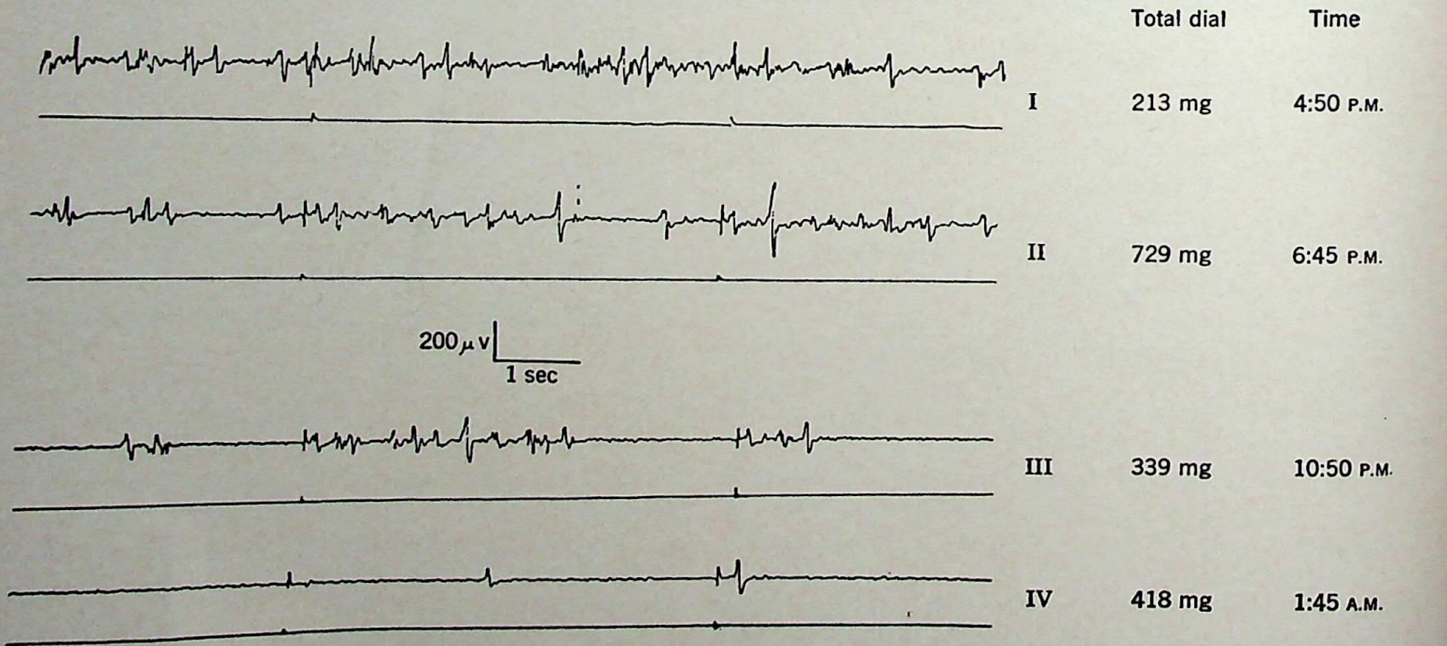


FIG. 7. Four tracings (I, II, III, IV) of electrical activity by a gross electrode from the auditory cortex of a cat. These records were taken during the presentation of clicks. The delivery of the stimuli is indicated on the accompanying lower traces. Depth of anesthesia increases from I to IV, as the amount of dial administered is increased.[32]

† Erulkar, Rose, and Davies[31] found, likewise, that about one-third of the units in the auditory cortex was unresponsive to acoustic stimulation.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

activity. The time has now come for solid-state recording devices that will permit the exploration of what might be called the "domain-structure" of sensory systems. It would, however, be foolish to expect that the data from 10 or 50 inputs could be handled adequately without the help of digital computers of appreciable capacity. One cannot be satisfied with visual inspection of records if one wants to find out whether unit X responds most of the time only if units Y and Z are inhibited or after units A, B, and C have fired in a certain order and separated by certain time intervals.

DETECTION OF EVOKED RESPONSES IMBEDDED IN BACKGROUND ACTIVITY

In recent years, as neurophysiologists attempted to study cortical responses to sensory stimuli in animals that were, physiologically speaking, in a "more normal" state, two problems arose: First, it became more difficult to detect the evoked responses that were now imbedded in a much livelier background activity of the cortex than had been observed in the previously used, deeply anesthetized preparations. Second, it became important

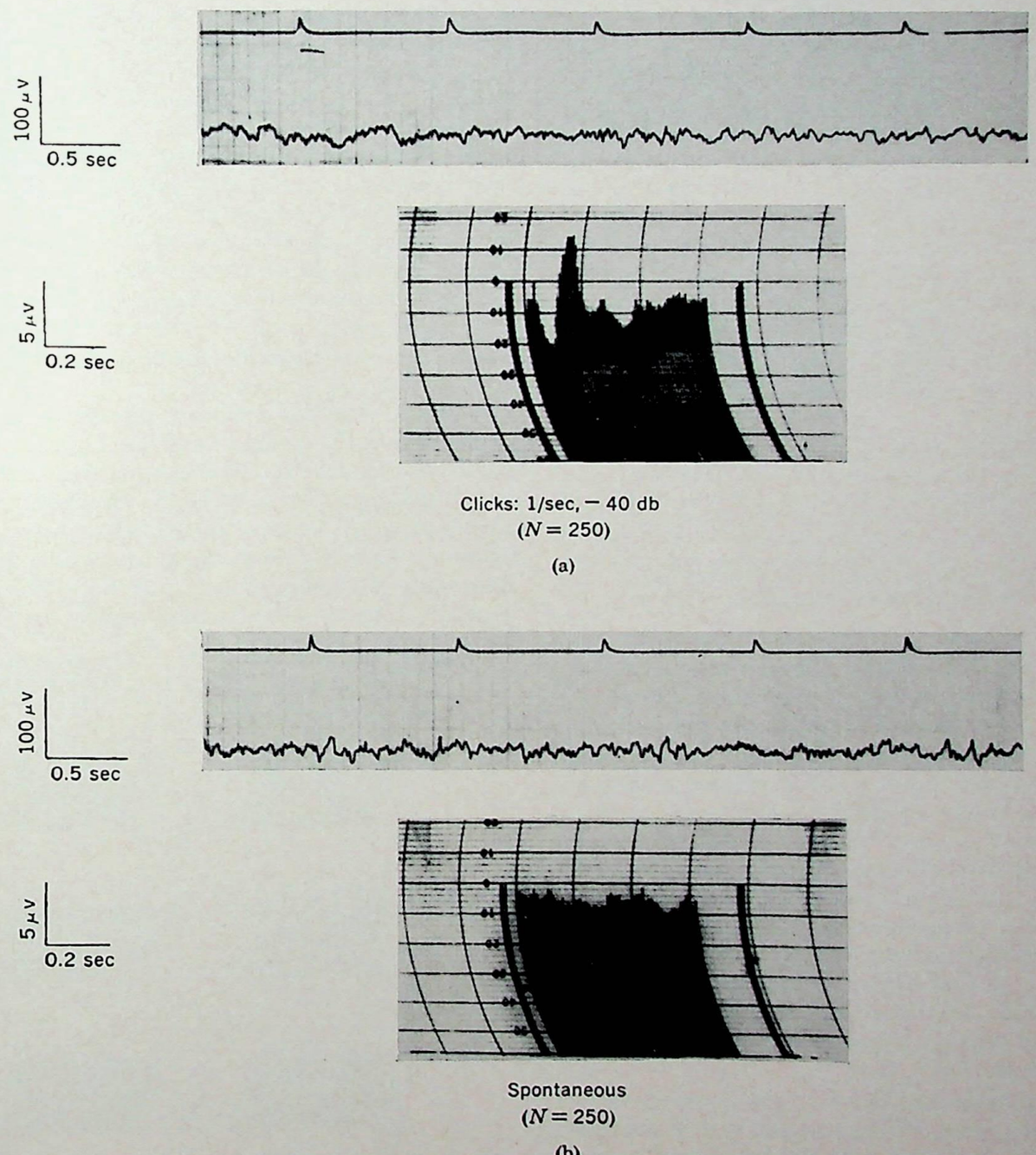


FIG. 8. The upper records of (a) and (b) show ink traces obtained from the skull (vertex) of human subjects (a) in the presence of clicks and (b) in the absence of clicks. The pulse channel, recorded simultaneously, indicates in (a) the instants at which clicks were delivered and in (b) serves as a comparable time reference. The lower records are the average wave forms obtained, by a process of cross correlation of the signal with a series of brief pulses,[14] from a sample of 250 intervals of which 5+ are shown in the ink records directly above. The average wave forms are the envelope of a series of closely spaced lines (10 msec).[33]

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

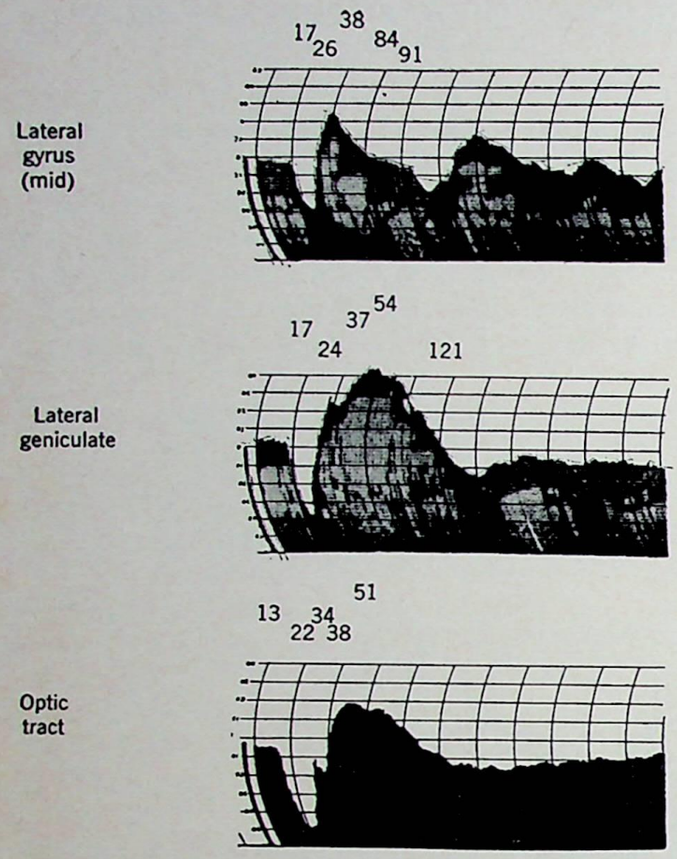


FIG. 9. Averaged responses to flash from three stations in the visual pathway of a cat; simultaneous recordings under moderate pentobarbital anesthesia with reference lead on the back of the neck. The average response is the envelope of a series of pen deflections separated by intervals of 1 msec. The delivery of the flash coincides with the first of the continuous series of pen deflections. The numbers above each average response are the latencies of characteristic points in the particular wave form [from M. A. B. Brazier, Acta Physiol. Pharmacol. Neerlandica 6, 692 (1957)].

to find ways of characterizing this background activity in a more than impressionistic manner.

Figure 7 illustrates the first problem by demonstrating how additional doses of anesthetic make it much easier to detect the "signal" (the evoked response) in the much diminished background activity that can, however, hardly be called "noise." Raab and Kiang[32] were actually looking for correlations between measures of the activity that preceded the delivery of a stimulus and the size of the evoked response. Their precise results are not of particular import here. It is, however, important to note that the interaction between evoked and background activity of the nervous system is not a problem for which a simple or even a single solution can be assumed.

Having a probabilistic bias in the interpretation of the electrical activity of the nervous system, members of our laboratory have attempted for several years to treat the obtained data in a manner that will yield average responses to sensory stimuli. The reasoning was, qualitatively, somewhat as follows: In the handling of sensory information, organisms behave as if they were acting on the basis of estimates of activity averaged over substantial regions of the nervous system; such a weighted averaging process could be carried out by comparing the outputs of a large number of neural elements. Since there are at present no devices capable of performing this task (even if one knew exactly where and how to perform it), one chooses to average over an ensemble of responses to repeated identical stimuli in order to bring out certain typical aspects of the behavior of the nervous system. Once the "correlator system" (referred to below) was available, it became possible to use it with comparatively slight modifications for the detection of evoked responses.[14] Figure 8 illustrates how this averaged display yields average results that would be hard to foretell on the basis of visual inspection either of ink tracings or of tracings of single responses photographed from the screen of an oscilloscope. Electronic computation of the sort used in our laboratory necessitates the recording of data onto FM magnetic tapes. Such tapes not only contain, in general, the specific responses that are set off by delivery of the stimuli but also offer, especially when multichannel recording is used, a much more representative sample of the organism's electrical activity. Appropriate programming permits the carrying out of different analyses of the recorded data as desired by replaying the tapes as often as is necessary.

Figures 9 (from Brazier[34]) and 10 offer just two examples of how the Evoked Response Detector (ERD) modification of the correlator system was used in the study of the visual[34] and auditory[35] systems. Other problems investigated include: cortical off-responses in the auditory system,[36] the effect of anesthesia upon the wave form of evoked cortical responses,[37] as well as responses to photic stimulation in man.[38]

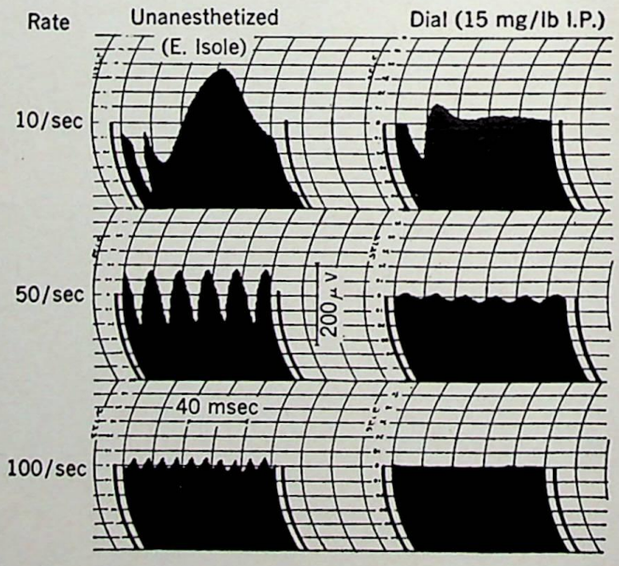


FIG. 10. Averaged cortical responses to repeated clicks (25-db threshold) indicate that the auditory cortex in nonanesthetized preparations is capable of "following" at higher rates. Note, also, the difference in wave form of response in anesthetized and nonanesthetized preparations for 10 clicks/sec. The evoked-response detector averaged the following number of responses at the different rates: 600 responses at 10/sec, 3000 at 50/sec, and 6000 at 100/sec.[35]

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

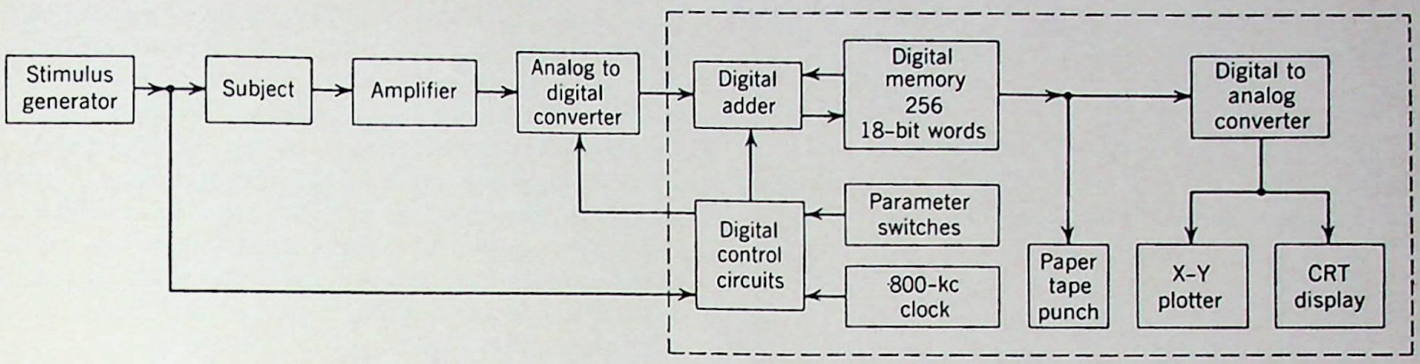


FIG. 11(a) Block diagram of average response computer (ARC-1).

More recently, we have been able to use (thanks to the cooperation of some colleagues at the Lincoln Laboratory) a considerably more powerful special-purpose digital computer, the ARC-1 (Average Response Computer[39]) in these investigations. Perhaps the most significant feature of this transistorized computer is that it computes "in real time" (Fig. 11). Thus, a display can be obtained instantaneously of the average of all response traces including the one just recorded from the animal's or subject's head. This immediate access to knowledge of results offers advantages, especially when as complex a system as the nervous system is being investigated. Without entering into the realm of science fiction, one can foresee experiments in the near future in which an organism and a computer operate in a "closed loop," i.e., in which the organism's response behavior alters the programming of sensory stimuli that the organism is instructed to deal with. With the help of ARC-1, our laboratory has continued (see Fig. 8) a rather exciting study of evoked potentials in man. It has thus been possible not only to record extracranially what appear to be cortical responses to acoustic clicks in many experimental subjects,[40] but also to demonstrate that these electrical responses appear at stimulus intensities that are quite comparable to the subject's absolute psychophysical threshold for clicks (see Fig. 12). Furthermore, it has been noted that the responses disappear near the subject's masked threshold. These findings suggest numerous sensory experiments in which convergent neurophysiological and psychophysical data can be obtained from human beings and are thus in tune with an era of experimentation in which the electrical activity of behaving animals will be recorded by implanted electrodes and broadcast by transistorized transmitters.[41]

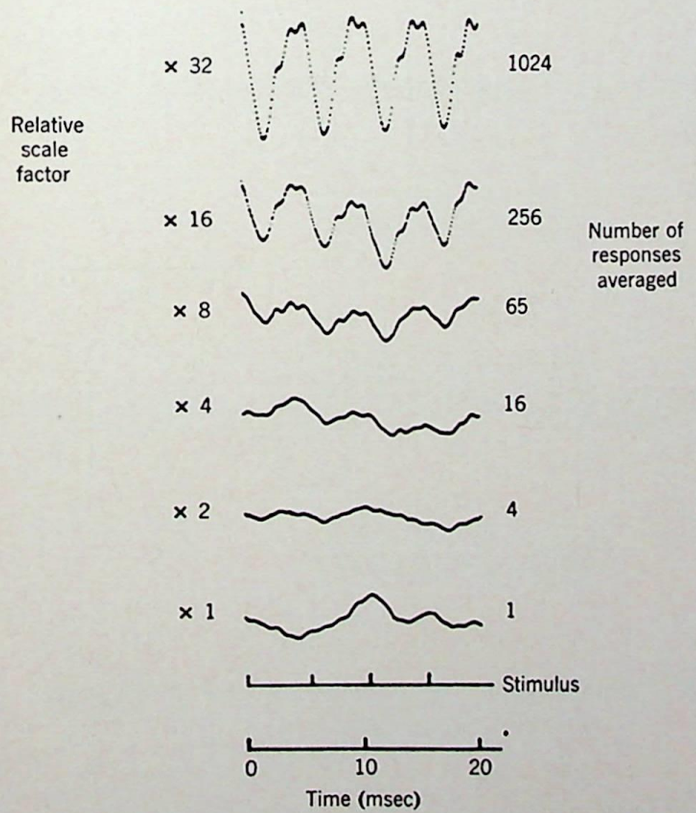


FIG. 11(b). Emergence of an average response from background activity as the number of averaged responses is increased.[39]

ATTEMPTS AT QUANTIFICATION OF ONGOING ACTIVITY

Up to this point, this paper has emphasized specific electrical responses to specific sensory stimuli, although it has been pointed out that these responses are, at all levels of the nervous system, in a nontrivial sense functionally related to the existing background activity.§

Trained electroencephalographers are clearly capable of arriving at diagnostic judgments on the basis of ink tracings of the EEG. Attempts have been made to "quantify" (here the term is used in its broadest meaning) various aspects of brain-wave activity in experimental situations. These attempts at quantification have ranged from a classification of the recorded complex visual patterns into broad classes (corresponding either to characteristics of the EEG wave form or to such states as *arousal* or *drowsiness*, or to some pathological condition) to a running Fourier analysis of short-time (several seconds) samples of the EEG activity. Given our lack of theoretical understanding of the physiological mechanisms that underlie the EEG[42] and the multiplicity of purposes for which EEG data have been used, it is not surprising that no ideal method

§ This background activity has been designated as *spontaneous* or *ongoing* activity. The precise conditions under which it is observed in pure form are extremely hard to define.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

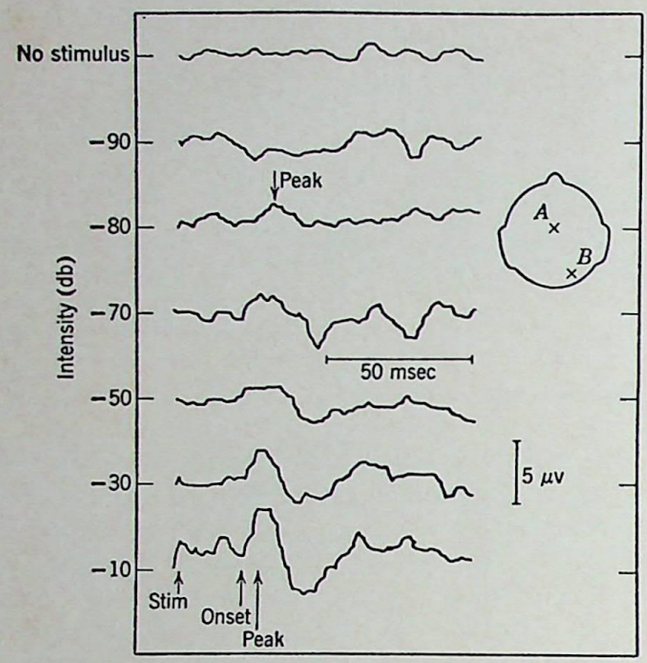


FIG. 12. Computed average responses to monaural periodic clicks obtained from scalp electrodes for different stimulus intensities. Each trace represents the wave form of the average response to 400 individual presentations of identical click stimuli delivered at a rate of 1.5/sec. The subject's (W. P., 10 September 1957; awake, eyes open) psychophysical threshold for these clicks, as determined during the same experiment, lay at approximately −85 db. Upward deflection indicates that electrode *A* is positive with respect to electrode *B* [from C. D. Geisler, L. S. Frishkopf, and W. A. Rosenblith, Science 128, 1210 (1958)].

of analysis has been found. Several years ago Wiener—intent on testing certain of his ideas regarding the presence of a "clock" in the brain[43]—interested Brazier's group at the Massachusetts General Hospital and the Communications Biophysics group in the Research Laboratory of Electronics, MIT, in applying correlation analysis to the EEG. A "correlator system"[44] was built, and a great quantity of experimental data has by now been processed.[45] There developed a considerable amount of discussion regarding the mathematical nature of the EEG time series, regarding the usefulness of correlation analysis of the EEG (since it discards so much of the information contained in the raw data), and even regarding the ease with which a correlation function of the EEG can be interpreted in a really quantitative manner. Whatever the merits of these issues, any data-reducing scheme should prove of some usefulness if the final display of the dependent variable satisfies several requirements: (a) it should be stable for a given organism when data are taken under comparable conditions; (b) it should be sensitive to important changes in the organism's environment; (c) it should perhaps be possible to classify a whole population on the basis of a limited number of displays.

Figures 13 to 15 illustrate some of the properties of the autocorrelation function as such a display. Figure 13 demonstrates that the display is capable of reflecting a threefold classification of the EEG ink tracings into

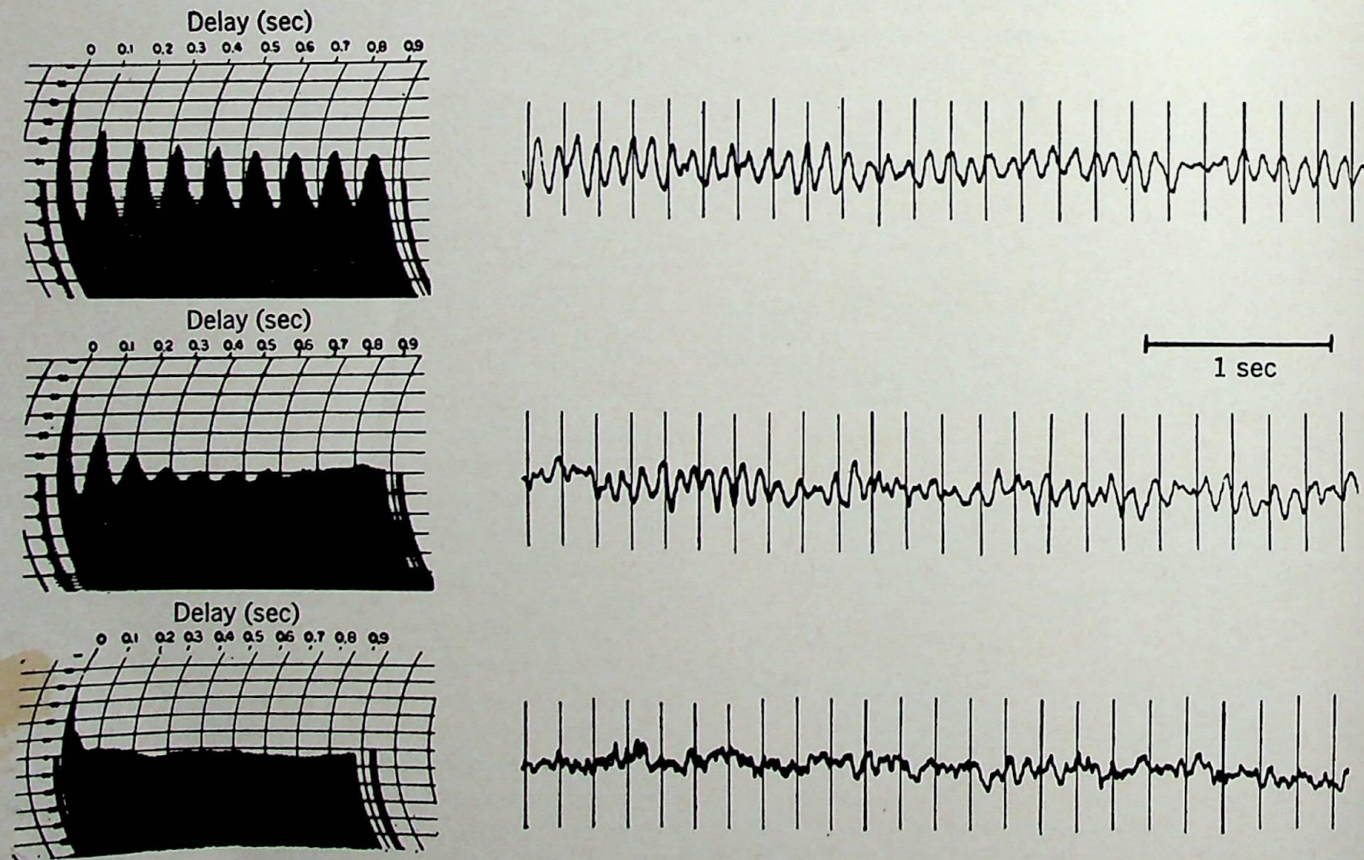


FIG. 13. Left: Autocorrelograms of 1-min samples of the EEG of three normal human subjects recorded from bipolar electrodes (right parietal to right occipital) on the scalp. Right: Sample of ink records from these same subjects [from J. S. Barlow, M. A. B. Brazier, and W. A. Rosenblith, "The application of autocorrelation analysis to electroencephalography," in *Proceedings of the National Biophysics Conference* (Yale University Press, New Haven, to be published)].

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

"high alpha," "medium alpha," and "no alpha." It remains to be seen to what extent a much larger population could be accurately classified on the basis of a certain number of such displays for various locations on the skull.

Figure 14 asks the question: How stable is such a display for a given person over a prolonged period of time? Although it is not possible at present to evaluate the similarity of these two displays in precise numerical terms, they appear to be quite comparable if the conditions of recording the EEG are reasonably well con-

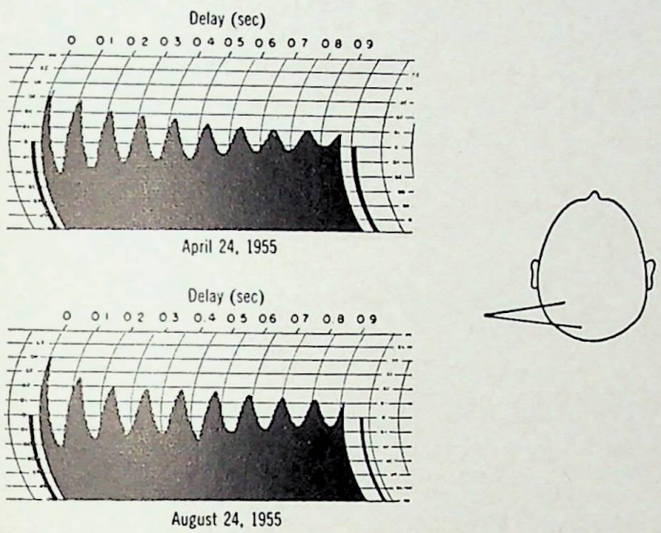


FIG. 14. Autocorrelograms of the EEG (1-min sample) of a normal human subject recorded four months apart [from J. S. Barlow, M. A. B. Brazier, and W. A. Rosenblith, "The application of autocorrelation analysis to electroencephalography," in *Proceedings in the National Biophysics Conference, 1957* (Yale University Press, New Haven, to be published)].

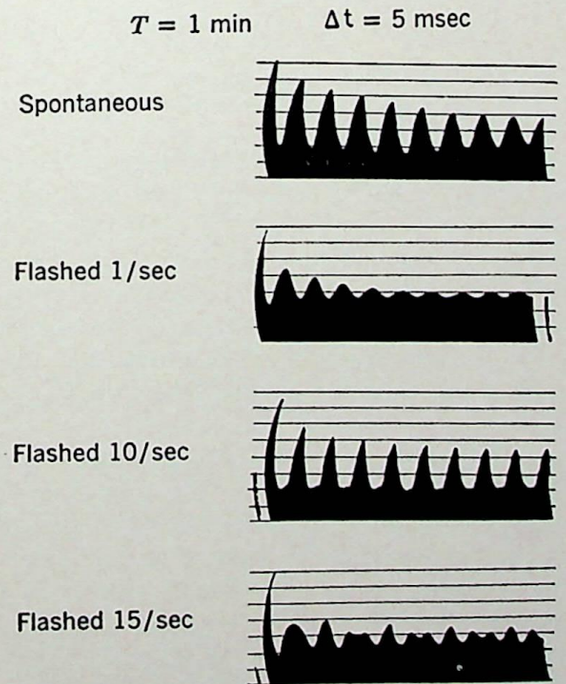


FIG. 15. Autocorrelograms computed for same subject in absence of controlled sensory stimulation ("spontaneous") and in presence of different rates of periodic photic stimulation [from J. S. Barlow, M. A. B. Brazier, and W. A. Rosenblith, "The application of autocorrelation analysis to electroencephalography," in *Proceedings of the National Biophysics Conference, 1957* (Yale University Press, New Haven, to be published)].

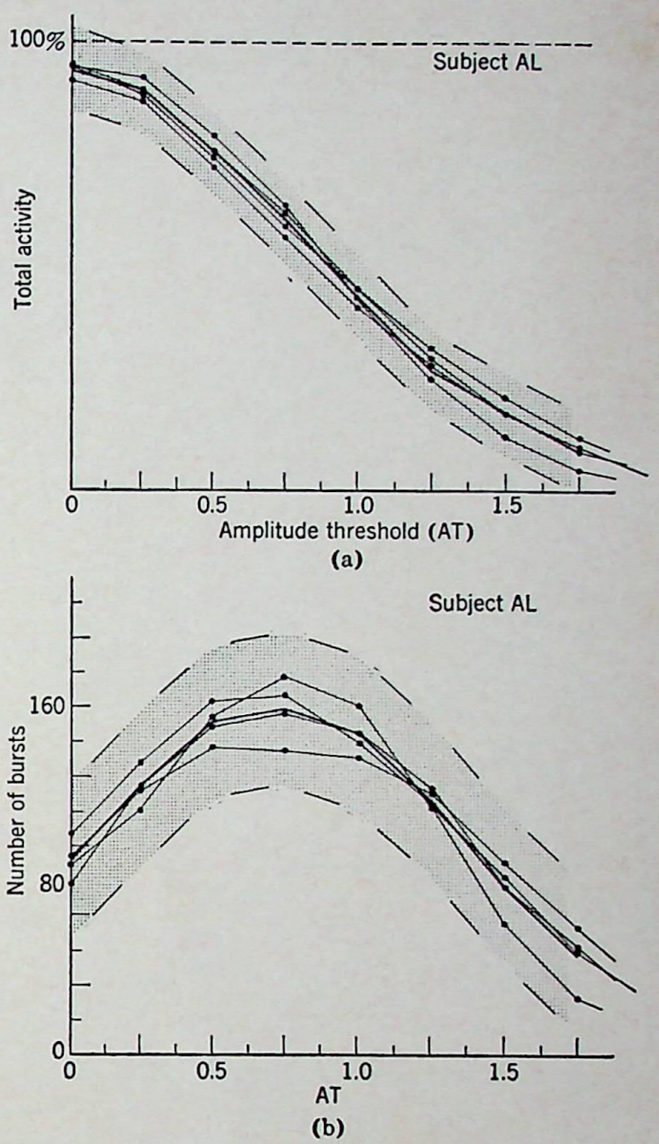


FIG. 16. Brain waves were recorded on four consecutive days (at same time of the day) from an awake subject sitting with closed eyes in an anechoic chamber. The displays show the four sets of results in fine lines, with the mean of the four runs as a heavy solid line. Each analyzed run lasted 3 min. For details of computer program and definitions of such terms as *amplitude threshold*, *total activity*, and *number of bursts*, see Farley *et al.*[46] The shaded region is bounded by the estimated confidence limits for the variation of the displays.[46]

trolled and the time sample from which the correlation function is computed is sufficiently long. Figure 15 demonstrates that the autocorrelational display is sensitive to photic stimulation.

It is not possible to examine here the extent to which the findings obtained with the aid of correlation (both autocorrelation and crosscorrelation) analysis support a particular model of brain function. One must, of course, admit that any method of data analysis implies a model of the process that is being investigated. The research worker has a burden of proof in the sense that he must show that the implied model fits the process well enough for his purpose. The fact remains that, in the absence of systematic computational analysis, most questions regarding the mathematical nature of the signal cannot even be raised in a meaningful manner. In addition, one should not underestimate the heuristic

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

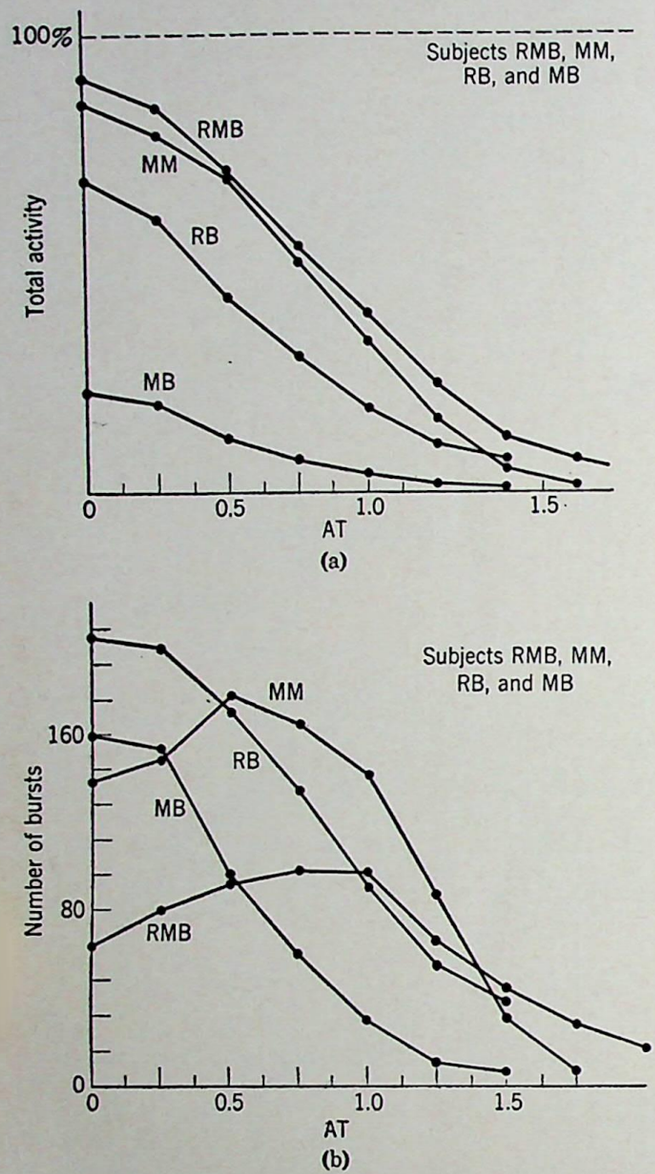


FIG. 17. Functions obtained for four different subjects on the basis of single 3-min runs. Note that subjects RMB and MM do not differ much in their total activity profile, but they are clearly distinguished in the display depicting the number of bursts.[46]

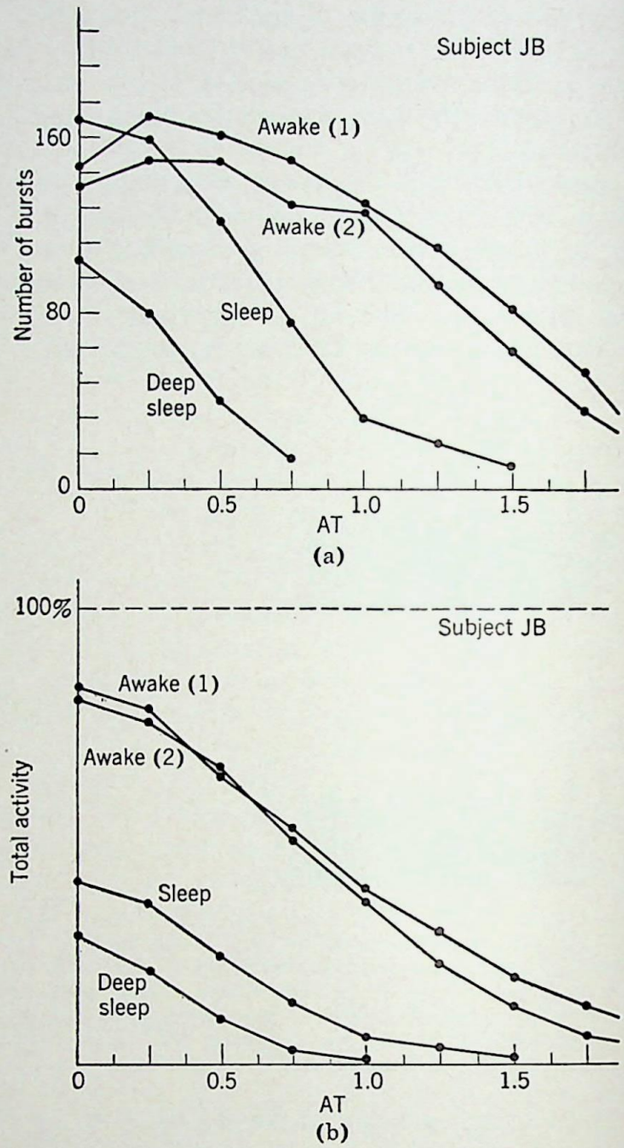


FIG. 18. Four runs for a subject, two while he was awake and two in two different states of sleep.[46]

benefits that derive from the use of a display that is in itself "anschaulich."

Displays of correlation functions of ongoing electrical activity are just one way of trying to describe an organism's "state" in terms of pseudo-state-variables. A report entitled "Computer Techniques for the Study of Patterns in the Electroencephalogram"[46] represents perhaps the most far-reaching effort in this direction. With the help of a rather flexible general-purpose digital computer equipped with a large memory, the attempt was made to "recognize the pattern of rhythmic bursts" in a given individual's EEG. Once a program for this purpose had been written, the attempt was made to distinguish different subjects from each other, as well as different "states" in the same subject.

Figures 16 to 18 show some of the results that have been obtained. Figure 16 illustrates that it is possible to obtain quite reproducible displays for the same subject if precautions are taken to obtain the data under fairly constant conditions. It was even possible to estimate confidence limits for the variability of the display for this subject and to compare displays from other subjects with these limits in order to assess significant differences. The initial results by Farley *et al.*[46] indicate that even a limited number of displays is capable of separating subjects into statistically distinguishable classes. Increasing the number of displays by increasing the number of pseudo-state-variables should make it possible to remove existing "degeneracies."

Figure 18 shows how these displays in a given subject are affected by his going to sleep. The influence of drugs or of sensory stimulation could be studied in this manner. Some evidence has already been obtained that these displays are sensitive to even short periods of sensory deprivation.

It is, however, appropriate to utter here a caveat regarding this and other methods that aim at quantifying data on the electrical activity of the nervous system. The multivariate nature of the nervous system prevents one from taking too much stock in any given number or even in a computed curve. Somehow one must learn to derive quantities that characterize the essentials of

CC-0. Library of Congress Collection. Digitized by Sarayu Foundation USA

neural activity in a fashion that is perhaps somewhat analogous to the dimensionless parameters of engineering systems. To make this point more concrete: Farley *et al.*[46] have shown that there are individuals whose total activity profile when they are awake bears some resemblance to the profile of another individual while the latter is asleep. However, whenever a person falls asleep there are changes in his activity profile. In other words, one must learn to assess the significance of change in relation to certain base lines and must refrain from attaching too much value to isolated pieces of numerical information. Unfortunately, economical mathematical descriptions of patterns and sequences are not yet available. These considerations can perhaps be carried somewhat further by noting that Fig. 13 illustrated how widely the autocorrelogram from apparently normal individuals may vary. It would thus seem unwise to draw too far-reaching conclusions regarding, for instance, a person's psychological make-up on the basis of a single display or even a set of summary displays computed from the electroencephalogram.

ACKNOWLEDGMENTS

The author owes a debt of gratitude to his colleagues and associates in MIT's Research Laboratory of Electronics. In connection with the data presented in this paper, he is particularly indebted to J. S. Barlow, M. A. B. Brazier, R. M. Brown, B. G. Farley, M. Z. Freeman, L. S. Frishkopf, C. D. Geisler, M. H. Goldstein, Jr., N. Y.-S. Kiang, J. Macy, Jr., C. E. Molnar, W. T. Peake, D. H. Raab, T. T. Sandel, and T. F. Weiss. To some extent, their views are reflected in the discussion, although the author must bear sole responsibility for the conclusions that he has drawn.

BIBLIOGRAPHY

[1] K. L. Chow, J. Comp. Neurol. **95**, 159 (1951).
[2] R. Galambos, J. Neurophysiol. **19**, 424 (1956).
[3] H. Davis, Physiol. Rev. **37**, 1 (1957).
[4] H. W. Magoun, *The Waking Brain* (Charles C. Thomas, Springfield, Illinois, 1958).
[5] G. Moruzzi and H. W. Magoun, Electroencephalog. and Clin. Neurophysiol. **1**, 455 (1949).
[6] E. D. Adrian, F. Bremer, and H. H. Jasper, editors, *A C.I.O.M.S. Symposium: Brain Mechanisms and Consciousness* (Blackwell Scientific Publications, Oxford, 1954).
[7] R. Hernández-Peón, H. Scherrer, and M. Jouvet, Science **123**, 331 (1956).
[8] S. W. Kuffler and C. C. Hunt, Research Publ. Assoc. Nerv. Mental Dis. **30**, 24 (1952).
[9] K.-E. Hagbarth and D. I. B. Kerr, J. Neurophysiol. **17**, 295 (1954).
[10] D. I. B. Kerr and K.-E. Hagbarth, J. Neurophysiol. **18**, 362 (1955).
[11] R. Granit, *Receptors and Sensory Perception* (Yale University Press, New Haven, 1955).
[12] R. Galambos, G. Sheatz, and V. G. Vernier, Science **123**, 376 (1956).
[13] M. Jouvet and J.-E. Desmedt, Compt. rend. **243**, 1916 (1956).
[14] J. S. Barlow, Electroencephalog. and Clin. Neurophysiol. **9**, 340 (1957).
[15] W. J. McGill, "A statistical description of neural responses to clicks recorded at the round window of the cat," Ph.D. thesis, Harvard University (1952).
[16] J. Macy, Jr., "A probability model for cortical responses to successive auditory clicks," Ph.D. thesis, Massachusetts Institute of Technology (1954).
[17] L. S. Frishkopf, Tech. Rept. 307, Research Laboratory of Electronics, Massachusetts Institute of Technology (1956).
[18] K. S. Harris, "Serial characteristics of neural responses to acoustic clicks recorded at the auditory cortex," Ph.D. thesis, Harvard University (1954).
[19] J. E. Rose and V. B. Mountcastle, Bull. Johns Hopkins Hosp. **94**, 238 (1954).
[20] M. H. Goldstein, Jr. and N. Y.-S. Kiang, J. Acoust. Soc. Am. **30**, 107 (1958).
[21] E. A. Blair and J. Erlanger, Am. J. Physiol. **106**, 524 (1933).
[22] C. Pecher, Arch. Int. Physiol. **49**, 129 (1939).
[23] W. A. Rosenblith, J. Acoust. Soc. Am. **22**, 792 (1950).
[24] W. A. Rosenblith in *Proceedings of the Symposium on Information Networks* (Polytechnical Institute of Brooklyn, 1954), p. 223.
[25] L. S. Frishkopf and W. A. Rosenblith in *Symposium on Information Theory in Biology* (Pergamon Press, New York, 1958), p. 153.
[26] V. E. Amassian and H. J. Waller in *Reticular Formation of the Brain* (Little, Brown and Company, Boston, 1958), p. 69.
[27] M. Palestini, G. F. Rossi, and A. Zanchetti, Arch. Ital. Biol. **95**, 97 (1957).
[28] G. F. Rossi and A. Zanchetti, Arch. Ital. Biol. **95**, 199 (1957).
[29] O. J. Grüsser and A. Grützner, v. Graefe's Arch. Ophthalmol. **160**, 65 (1958).
[30] R. Jung, O. Creutzfeldt, and O. J. Grüsser, Deut. med. Wochschr. **82**, 1050 (1957).
[31] S. D. Erulkar, J. E. Rose, and P. W. Davies, Bull. Johns Hopkins Hosp. **99**, 55 (1956).
[32] D. H. Raab and N. Y.-S. Kiang, Quart. Progr. Rept., Research Laboratory of Electronics, Massachusetts Institute of Technology (October 15, 1955), p. 76.
[33] L. S. Frishkopf and C. D. Geisler, Quart. Progr. Rept., Research Laboratory of Electronics, Massachusetts Institute of Technology (January 15, 1957), p. 123.
[34] M. A. B. Brazier, Acta Physiol. Pharmacol. Neerlandica **6**, 692 (1957).
[35] M. H. Goldstein, Jr., N. Y.-S. Kiang, and R. M. Brown, J. Acoust. Soc. Am. **31**, 356 (1959).
[36] N. Y.-S. Kiang and T. T. Sandel, Quart. Progr. Rept., Research Laboratory of Electronics, Massachusetts Institute of Technology (July 15, 1957), p. 139.
[37] N. Y.-S. Kiang and M. H. Goldstein, Jr., Quart. Progr. Rept., Research Laboratory of Electronics, Massachusetts Institute of Technology (July 15, 1957), p. 142.
[38] M. A. B. Brazier in *Reticular Formation of the Brain*, H. H. Jasper *et al.*, editors (Little, Brown and Company, Boston, 1958), p. 151.
[39] W. A. Clark, Jr., Quart. Progr. Rept., Research Laboratory of Electronics, Massachusetts Institute of Technology (April 15, 1958), p. 114.
[40] C. D. Geisler, L. S. Frishkopf, and W. A. Rosenblith, Science **128**, 1210 (1958).
[41] G. E. Forsen, "An FM transmitter for recording the electrical activity of the nervous system," S.M. thesis, Massachusetts Institute of Technology (June, 1957).
[42] M. A. B. Brazier, J. Nerv. Mental Dis. **126**, 303 (1958).
[43] N. Wiener, Scientia **52**, 1 (September, 1958).
[44] J. S. Barlow and R. M. Brown, Tech. Rept. 300, Research Laboratory of Electronics, Massachusetts Institute of Technology (1955).
[45] J. S. Barlow, "Autocorrelation and crosscorrelation analysis in electroencephalography," Eleventh Annual Conference on Electrical Techniques in Medicine and Biology," 19–21 November 1958, Minneapolis (manuscript in preparation).
[46] B. G. Farley, L. S. Frishkopf, W. A. Clark, Jr., and J. T. Gilmore, Jr., Tech. Rept. 337, Research Laboratory of Electronics, Massachusetts Institute of Technology, and Tech. Rept. 165, Lincoln Laboratory, Massachusetts Institute of Technology (1957).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

58

Imitation of Pattern Recognition and Trial-and-Error Learning in a Conditional Probability Computer*

ALBERT M. UTTLEY

National Physical Laboratory, Teddington, Middlesex, England

A CONDITIONAL probability computer, appropriately connected with input and output devices, can "recognize" patterns and "learn" by trial and error. The principle of such a system developed at the National Physical Laboratory is described briefly in this paper, together with some comments on the role of engineering in the study of biological problems.

An engineer generally starts with a specification, a set of rules of behavior for a machine, and then tries to design a machine that possesses that behavior. He designs an airplane, for example, to fly at certain speeds, to go certain distances, and to carry certain loads; from these requirements he deduces the required structure.

Nature is the engineer that constructs the difficult bridge between structure and behavior of living systems. In psychology, one discusses the behavior of an organism and wonders what kind of structure could cause it. In physiology and anatomy, on the other hand, one usually studies the structure and wonders what it is doing, what function it is supposed to serve.

We believe, in all humility, that the engineer can perhaps help to bridge the gap between structure and behavioral function in biological systems. Using his known methods, the engineer can most fruitfully start with observed behavior and work toward a possible elucidation of structure. Some engineers have tried to work in the opposite direction: to think up some properties of units, give them random connections, and find out what they will do. We do not think that that approach has been fruitful. We have considered, specifically, the problems of pattern recognition and learning, and have tried to design computers to do these things in the hope that the structure of the computers might be suggestive of the organization of the nervous system.[1]

This approach differs in one respect from that of Shannon and many others, who have also studied the design of machines to do certain things which animals and men can do. Their approach has been to program a universal computer. Such an approach should lead to a deep mathematical understanding of behavior, but it is not likely to lead to a specific structure which will give clues to the physiologist, because a universal machine can do anything.

The first step toward synthesizing pattern recognition is to devise a method for distinguishing one *set* of signals from another, corresponding to sets of signals in nerve fibers. The signals might be continuously variable, or they might be discrete and binary; the latter type is much easier to handle, and is used in this study. The machine must have a component unit for each signal-input channel to indicate if that channel is active; and a unit for each possible pair of channels to indicate, by means of a coincidence circuit, whether or not that pair of channels is active. In general, one needs a unit for each possible set of channels. Such an arrangement is called a classification system; an example is shown in Fig. 1.[2]

Next, the system must be made to recognize patterns in *time*; this requires the introduction of delay circuits. Suppose that there are only two input channels, and that each is put through a fixed delay. If these two inputs and their delayed versions are used as four inputs of a classification system, the system will recognize all 12 "tunes" which can be played on a two-note piano if one allows only two temporal ideas: "before" and "after." The connections and recognition possibilities of such a system are shown in Fig. 2.

A classification system can be built with quite *random* connections. The reason is that all possible arrangements of connections between units and channels are required. If there are sufficient units to accommodate the desired inputs, and if they are connected entirely at random to the inputs, a classification system will arise automatically.

Next, the system must be given *plastic* behavior, in which past events can influence future responses. This property is introduced through the computation of *conditional probability*. Suppose that there are two

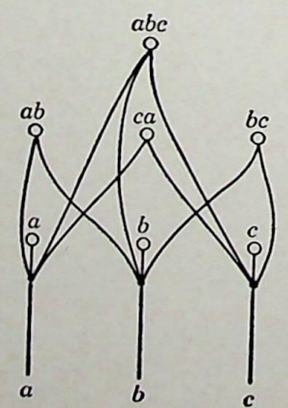


FIG. 1. A classification system for three inputs has seven units.

* This paper is published by permission of the Director of the National Physical Laboratory.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

events, A and B, which occurred as follows on nine occasions:

A: $I\ I\ I\ I\ I\ O\ O\ O\ O$ (I = occurrence,
B: $O\ O\ I\ I\ I\ I\ I\ I\ I$ O = nonoccurrence)

One now asks: "If the future is like the past—if one accepts the inductive principle—and if A occurs, what is the chance of B?" In the past, A occurred on five occasions, and on three of these B also occurred. The corresponding probability of B, given A, is three out of five. Conversely, the probability of A, given B, is three out of seven. A machine to make these calculations must be able to count the five, in an A unit; the seven, in a B unit; and the three, in a coincidence-counting AB unit. Thus, to possess plastic behavior, the machine needs an additional feature: each unit must be able to *count*, even if in a very nonlinear way. The system will then be capable of inductive inference.[3]

Another important matter which comes easily out of this approach is the "context" in which events occur. Suppose that, in the foregoing example, an event C occurred on the last five occasions:

A: $I\ I\ I\ I\ I\ O\ O\ O\ O$
B: $O\ O\ I\ I\ I\ I\ I\ I\ I$
C: $O\ O\ O\ O\ I\ I\ I\ I\ I$

This history now has two different contexts (C occurs and C does not occur), and one can ask: "In the context C, what would be the probability of A, given B?" In this context, B occurred five times and A once, so the three-out-of-seven chance has changed to one-out-of-five. Similarly, the probability of B, given A, in the context C, is unity. The system can be extended indefinitely to infer one thing from another in all sorts of different contexts.[4]

There is a practical difficulty in the design of counters: after a time they get full. The engineer has to put a leak into the counter to avoid this dilemma—and a very interesting property is thereby produced. The leaky counter gives a time weighting in which past events matter less than recent ones. The computer can then *adapt* fairly quickly to new conditions.

Finally, the system must be able to *calculate ratios* of

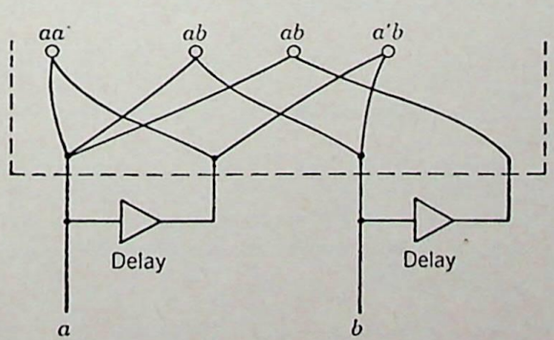


FIG. (2a). A classification system for 2 inputs which distinguishes "before" and "after." Four of the 15 units are shown.

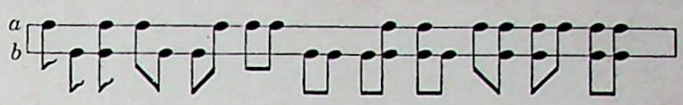


FIG. 2(b). The 12 tunes which can be distinguished by the system of Fig. 2(a). For the first 3 tunes—a, b, and ab—there are duplicate units.

numbers. Division is done most easily by using logarithms. It turns out that a leaky counter produces something very close to a logarithmic count, so the required ratios can be obtained simply by taking the differences between the states of two units. (The "state" is just the stored voltage in the counter.) Clearly, the effect of an active counting unit on one to which it is connected must depend on their relative state.

This relationship suggests a very important feature of a system that is capable of learning. The old idea that "facilitation" of a path depends on the number of times that it is used is not adequate; it is as illogical as the story of the Irishman and his boots—he couldn't get them on until he had worn them a few times. It is well known that a novel stimulus may, initially, produce no reaction from an animal, but eventually produce a response owing to trial-and-error learning. If initially a path is not made at all, how could it possibly be made on the principle that the more often it is used, the better it is?

The difficulty is resolved by a comparison of the states of two units. In neural terms, if a probability is to be computed, the numerator must be stored on one side of a synapse and the denominator on the other, so that a comparison can be made.

In summary, a machine to simulate pattern recognition and learning must be able to:

(a) distinguish sets of signals, through the use of coincidence counting;
(b) recognize time patterns, by use of delay circuits;
(c) count occurrences of signals and sets of signals;
(d) store information regarding past occurrences, at least for a certain length of time; and
(e) calculate ratios to determine relative states of units with stored counts of occurrences.[5]

These properties are incorporated in the machine shown in Fig. 3.

CONCLUDING REMARKS

Mathematical models can be very useful for studying complex systems and functions, including those of living organisms. The utility of a model, and especially the confidence with which one can draw inferences from its behavior, can be greatly enhanced by building a physical analog of the model—a machine. In the process of building the machine and making it work, one may find shortcomings or omissions in the theoretical model. More important, one may find that the machine is capable of doing more than was anticipated. The system reported here, for example, was designed to do pattern

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

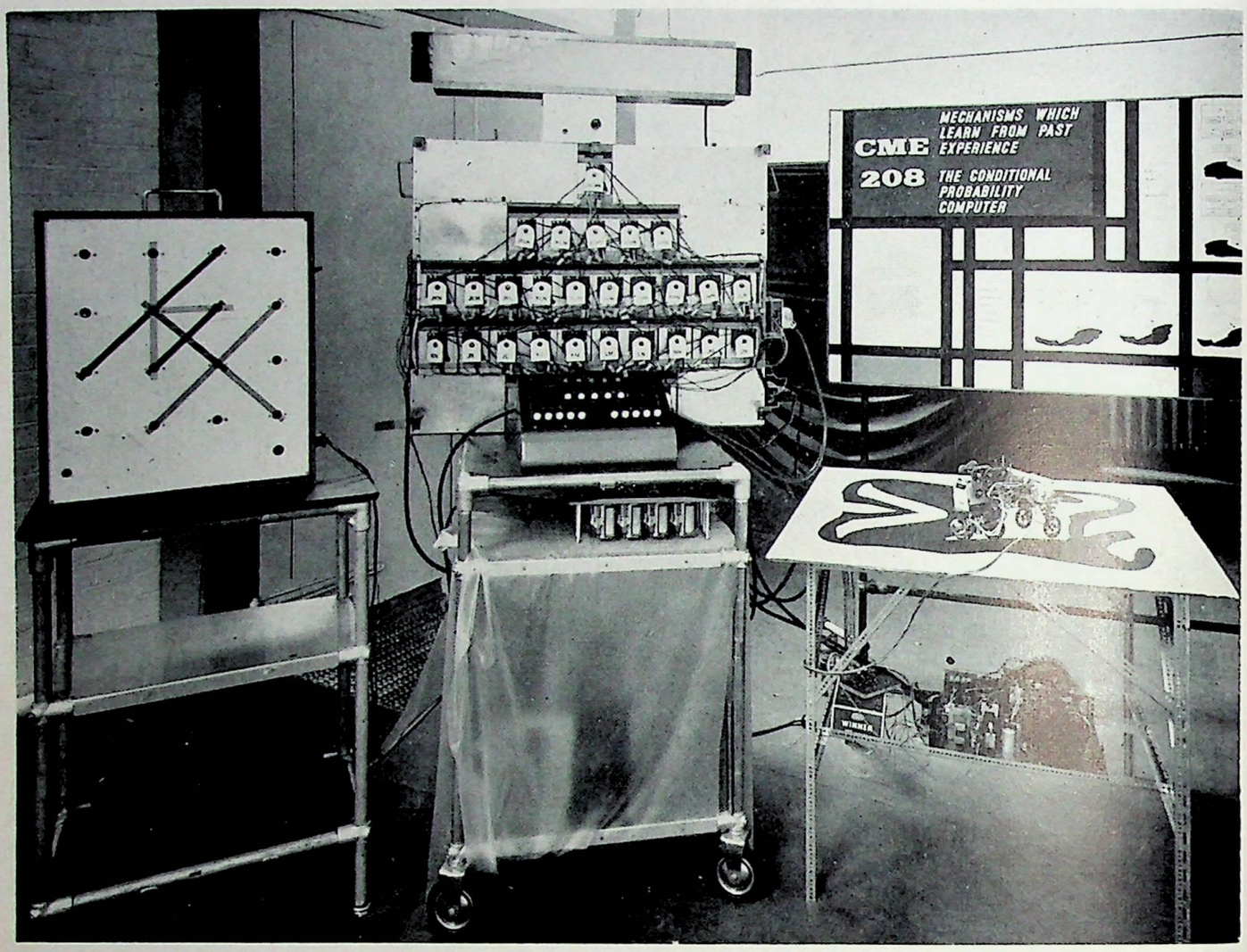


FIG. 3. A conditional probability computer, built at the National Physical Laboratory, capable of pattern recognition and trial-and-error learning. On the left is a model "retina" of photo cells to receive T-shaped shadow patterns. On the right is an electrically driven "vehicle" that learns, with the help of its computer "brain" which is shown in the center of the picture, to move efficiently along the black-white border of its "highway."

discrimination, and it turned out to be capable of trial-and-error learning as well.

Another important reason for making models is that people in different specialties use different languages—as is amply evident in a conference such as this Study Program. People try to talk about the same thing and hardly understand one another. A model is a sort of universal language. No matter what your field is, if you look at a physical model and see it working—see its parts working—you can get something out of it. In particular, the physiologist and anatomist can look at an engineer's model of a psychologist's observation and consider whether the structure of the model bears any resemblance to the structure of the living organism.[6]

One final comment: engineers use a very limited number of types of vacuum tubes to do a very large number of different things. The nervous system appears to be similar. There must be an enormous number of things neurons can do with just a few basic properties. The computation of conditional probabilities is, perhaps, just one of them.

BIBLIOGRAPHY

[1] A. M. Uttley, *Conditional Probability as a Principle in Learning*, Namur Cybernetics Conference (1956) (Gauthier-Villars, Paris) (to be published).

[2] A. M. Uttley, Electroencephalog. and Clin. Neurophysiol. 6, 479 (1954).

[3] A. M. Uttley in *Automata Studies*, C. E. Shannon and J. McCarthy, editors, Ann. of Math. Studies 34, 253 (1956).

[4] A. M. Uttley, Electroencephalog. and Clin. Neurophysiol. 6, 277 (1954).

[5] A. M. Uttley, "The design of conditional probability computers," Information and Control (to be published).

[6] A. M. Uttley, *Conditional Probability in a Nervous System*, N. P. L. Symposium (1958) (Her British Majesty's Stationery Office, London) (to be published).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

REVIEWS OF MODERN PHYSICS VOLUME 31, NUMBER 2 APRIL, 1959

59

Chemical Specificity in Biological Systems

WALTER KAUZMANN

Department of Chemistry, Princeton University, Princeton, New Jersey

THE importance of specificity in biology has been mentioned or implied in a great many contributions presented in this volume, and many examples have been discussed. It is worthwhile, however, to examine the concept of specificity by itself, and it is advisable to review some of the more important physical and chemical factors that are currently believed to be involved in the phenomenon.

In this paper are considered some typical examples of biological specificity involving more or less well-defined chemical systems as one of the specific interacting agents. The molecular basis of specificity is discussed with special emphasis on the concept of complementarity. Finally, some words are devoted to a review of the important and old-fashioned idea of specificity as a form of organization.

Serological specificity was defined by Landsteiner[1] as "the disproportional action of a number of similar agents on a variety of related substrata," and this definition is readily extended to other fields in which biological specificity manifests itself. Boyd[2] points out that this definition "would also include many chemical reactions, as might be expected if serological specificity were fundamentally chemical in nature, which we now believe," and again there is no need to hesitate in applying the remark to biological specificity in general. Indeed, specificity is an important aspect of chemistry; qualitative and quantitative analysis in organic and inorganic chemistry depends in a large degree upon "disproportionate actions" of similar agents on "related substrates" (e.g., the precipitation of silver halides in water and the separation of silver chloride from the bromide and iodide by treatment with ammonia, H_2O and NH_3 being regarded as similar agents and the silver halides being related substrates).

It is important to emphasize at the outset, too, that biological specificity is very variable in degree. Certain systems show a wide range of specificity whereas other systems show extremely narrow selectivity. What is likely to impress the chemist about biological specificity is the sharpness with which closely similar substances can be distinguished in some instances, but this sharpness is not at all universal.

It is worth mentioning that the concept of specificity is involved in the actions of inhibitors as well as in those of normal substrates, since many closely related biological systems are poisoned to decidedly different degrees by substances having similar chemical compositions. The drug industry thrives on this fact—as does mankind in general.

EXAMPLES OF CHEMICAL SPECIFICITY IN BIOLOGICAL SYSTEMS

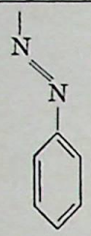
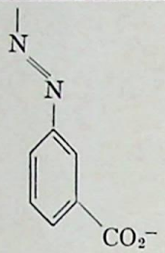
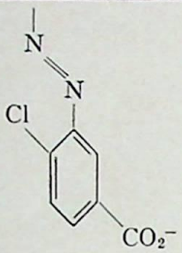
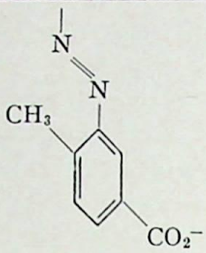
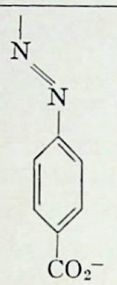
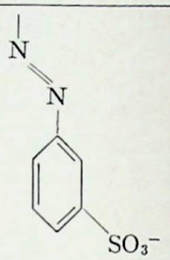
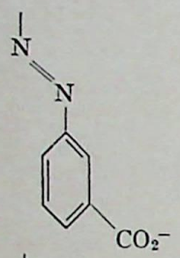
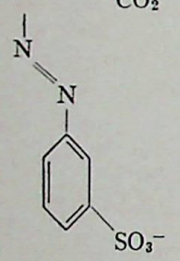
1. Antibody-Antigen Interactions

Because of the pioneering work of Landsteiner,[1] the field of immunochemistry provides one with probably the most easily understood examples of biological specificity, and it has made possible a fruitful method of attack on the relationship between specificity and chemical structure. When certain foreign substances (antigens) are introduced into an organism, the organism reacts by producing proteins (antibodies) which are capable of reacting specifically with the antigens that gave rise to them. Antigens are invariably substances of rather high molecular weight, so that it is difficult to alter their chemical composition in a systematic fashion. Landsteiner was able to show, however, that new antigens may be prepared by combining small chemical groups (such as substituted aromatic rings) with proteins. These new antigens give rise to antibodies whose specificity depends on the nature of the conjugated groups. The small chemical group is called a hapten. Let w, x, y, z represent a series of chemically related haptens, let Aw, Ax, Ay, Az represent antigens produced from protein A by conjugation with these haptens, and let Bw, Bx, By, Bz represent the antibodies arising from injection of Aw, Ax, Ay, Az into a rabbit. It is clear that a great deal might be learned about the chemical factors involved in immunological specificity by comparing the interactions of the different antibodies with the different antigens. Furthermore, it is found that a low molecular-weight compound, aw, containing the hapten group w, is able to inhibit the interaction between Aw and Bw. Evidently, Bw contains sites which are capable of reacting with the hapten groups on Aw, and the haptens in aw can compete for these sites. The study of this competition using a series of related haptens makes possible a more quantitative study of antibody-antigen specificity.

Some typical results of this general approach are reported in a paper by Landsteiner and Lampl.[3] Metaaminobenzoic acid and metasulfanilic acid were coupled by diazotization of the amino group to the proteins of horse serum, and antibodies to these modified proteins were obtained by injection into rabbits. The resulting rabbit antisera then were mixed with samples of chicken sera which had been diazotized to aniline or to various aniline derivatives. Evidence for interaction consisted in the formation of a precipitate. (The use of horse serum in the production of antibodies and of chicken serum in the antibody-antigen test

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

TABLE I. Effects of structural changes in the hapten on antibody-antigen interaction.[a]

Hapten present on horse-serum antigen used in preparing antisera.	Hapten present on chicken-serum antigens used to test interaction with antisera.					
						
	0	+++	++++	+++	0	+
	0	0		0	0	++++

[a] 0 = no precipitate; +, +++, ++++ = increasing amount precipitate.

eliminates any contribution of the nonhapten portion of the antigen to the antibody-antigen interaction.) The results are shown in Table I. It is seen that:

1. No interaction will occur if an acidic group is not present on the hapten of the chicken-serum antigen.
2. No interaction will occur unless the acidic group is meta to the amino group (i.e., CO_2^- and SO_3^- must be meta to the point of coupling of the hapten to the protein).
3. Replacement of a hydrogen atom on the benzene ring by a chlorine atom or a methyl group has little effect on the interaction.
4. The antibodies are readily able to distinguish between a carboxyl group and a sulfonic-acid group on the antigen.

On the basis of many experiments of this type, it could be concluded that the nature of the charged groups of the hapten is of decisive importance in determining the specificity. Haptens containing sulfonic-acid groups do not tend to compete effectively for sites on antibodies which are generated by carboxyl-containing haptens, and vice versa. The introduction of an arsenic-acid group into an antigen makes for a particularly strong interaction between the modified antigen and the antibodies which it induces. Introduction of an uncharged methyl, halogen, methoxyl, or nitro group into a hapten has a much smaller effect on the specificity; strong cross reactions are observed between antigens containing a given hapten and antibodies induced from antigens containing haptens that differ from the given hapten merely through the presence of one of these uncharged groups. On the other hand, antibodies are found to be able to distinguish very well between two haptens which are optical isomers or which are (*cis*-)(*trans*-) isomers.

The important inference to be drawn from these studies is that the specificity of the antibody-antigen reaction must depend upon a complementarity between the surfaces of the antibody and antigen molecules. This complementarity presumably involves an apposition of positive and negative electric charges, if a charge is present in the antigen. It also involves a rather close matching of the contours of portions of the surfaces of the antibody and antigen molecules—a lock and key relationship that had been visualized already by Ehrlich early in the development of immunology. Small deviations in the perfection of this complementarity are tolerated (e.g., replacement of a methyl group by a chlorine atom, or even of a hydrogen atom by a methyl group). Forces of all types are presumably in action across the surface of contact of the antibody and antigen, but electrostatic forces seem to be particularly important. One can visualize these relationships by means of semischematic drawings such as Fig. 1, where a portion of the antibody molecule

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

is supposed to envelop the hapten, with electric charges suitably located on the antibody surface so as to interact favorably with the opposite charge on the hapten.

A paper by Pressman, Siegel, and Hall[4] illustrates the possibility of obtaining quantitative information about complementarity by making use of the hapten-inhibition phenomenon. Samples of ovalbumin were diazotized to ortho-, meta-, and para-aminobenzoic acid. Three antigens were produced in this way, consisting of ovalbumin molecules bristling with azobenzoate groups having the three isomeric configurations shown in Fig. 2. The three antigens are referred to as "X_o-ovalbumin," "X_m-ovalbumin," and "X_p-ovalbumin." The three corresponding antibodies were prepared by injecting these antigens into rabbits, bleeding the rabbits, and separating the globulin fraction (which contains the antibodies) by ammonium-sulfate precipitation. The addition of suitable proportions of X_o-ovalbumin to anti-X_o-globulin gave a visible precipitate, the amount of which could be measured. When low molecular-weight haptens [such as benzoate, *o*-chlorobenzoate, *m*-chlorobenzoate, *p*-chlorobenzoate, *o*- (*p*′-hydroxyphenylazo) benzoate, and various chlorinated derivatives of *o*- (*p*′-hydroxyphenylazo) benzoate] were added, the amount of the precipitate was decreased because of the competition of the small haptens for the complementary site on the antibodies. By varying the concentration of low molecular-weight hapten and observing the change in the amount of precipitate produced, it was possible to compare the affinities of the different hapten groups for the sites on the anti-X_o-globulin.* These relative affinities could be expressed in terms of the differences between the mean free energies of combination of the haptens for the sites on the anti-X_o-globulin.

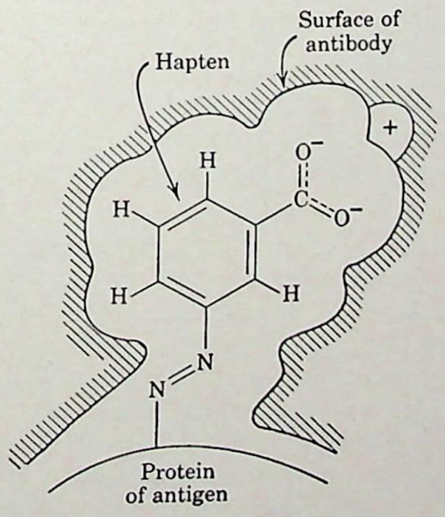


FIG. 1. Diagrammatic representation of complimentarity of the hapten-antibody relationship.

TABLE II. Relative affinities of various haptens for anti-$X_{o,m,p}$-globulins.

Hapten	Δ F_{rel}. cal/mole of hapten
For anti-X_o-globulin:	
Benzoate	(0)
o-chlorobenzoate	−580
m-chlorobenzoate	0
p-chlorobenzoate	430
o-(*p*′-hydroxyphenylazo) benzoate	−1700
3-Cl, 2-(*p*′-hydroxyphenylazo) benzoate	−1100
4-Cl, 2-(*p*′-hydroxyphenylazo) benzoate	−1100
5-Cl, 2-(*p*′-hydroxyphenylazo) benzoate	−1300
6-Cl, 2-(*p*′-hydroxyphenylazo) benzoate	−1500
For anti-X_m-globulin:	
Benzoate	(0)
o-chlorobenzoate	680
m-chlorobenzoate	−300
p-chlorobenzoate	480
m-(*p*′-hydroxyphenylazo) benzoate	−1600
2-Cl, 3-(*p*′-hydroxyphenylazo) benzoate	−560
4-Cl, 3-(*p*′-hydroxyphenylazo) benzoate	−350
5-Cl, 3-(*p*′-hydroxyphenylazo) benzoate	−980
6-Cl, 3-(*p*′-hydroxyphenylazo) benzoate	−500
For anti-X_p-globulin:	
Benzoate	(0)
o-chlorobenzoate	960
m-chlorobenzoate	−300
p-chlorobenzoate	−560
p-(*p*′-hydroxyphenylazo) benzoate	−1700
2-Cl, 4-(*p*′-hydroxyphenylazo) benzoate	−940

Similar experiments were performed on the X_m-ovalbumin-anti-X_m-globulin system and on the X_p-ovalbumin-anti-X_p-globulin system. As a result, it was possible to observe the effects that different groups in different positions in the hapten have on the ability of the hapten to combine with a given antibody. Some results for a series of chlorobenzoates and chloro-*p*-hydroxyphenylazobenzoate haptens are shown in Table II. The numerical values of ΔF_{rel} listed in this table are the differences between the average free energy for the combination of the hapten in question with the antibody sites and the average free energy of combination of the unsubstituted benzoate ion with the same sites. It is seen that, with each of the three globulins, the *p*-hydroxyphenylazo group greatly increases the affinity as compared with the benzoate ion (makes ΔF_{rel} more negative), presumably because it is able to take advantage of some of the binding forces associated with the portion of the antibody site that the organism intended to be occupied by the azo group and by the side chain of the protein to which the azo group is attached (see Fig. 3). This simple picture also shows

* The detailed mathematical analysis by Pauling, Pressman, and Grossberg[5] of the dependence of the amount of precipitate on the concentration of the low molecular-weight hapten showed that there must be a considerable variability in the affinities of the hapten for different sites on the antibody molecules. That is, the antibody sites at which the haptens are bound are not all alike. This heterogeneity can be described adequately by means of a Gaussian distribution of the free energies of binding about a mean value which is characteristic of each hapten, the spread in the free energies on either side of the mean value amounting to 1000 to 2000 cal/mole of hapten. The complication introduced by heterogeneity can be avoided by talking in terms of the mean free energy of binding for a given hapten at the sites on the anti-X_o-globulin.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

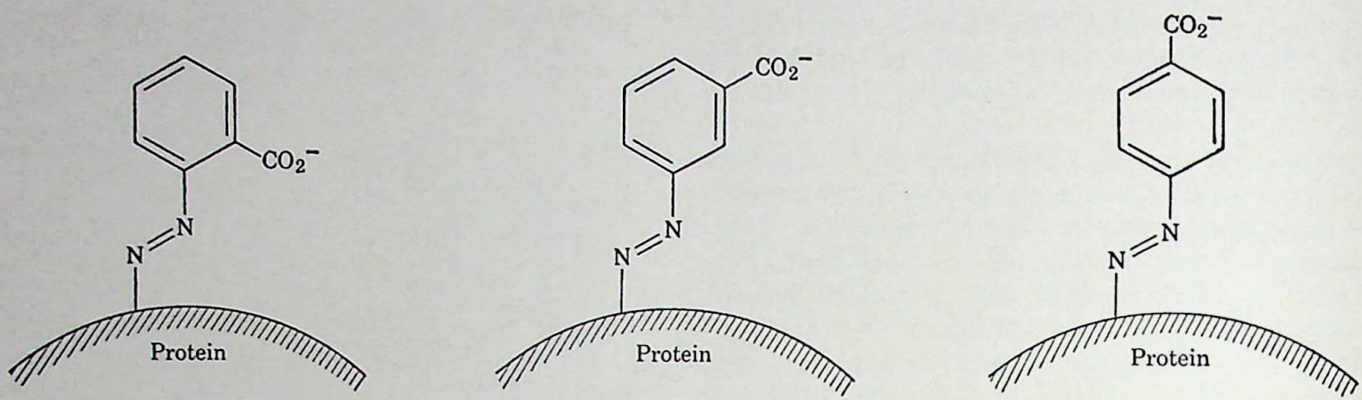


FIG. 2. Configuration of haptens used in studying antibody-antigen complementarity.

why the introduction of a chlorine atom in place of a hydrogen atom at various positions in the benzene ring in the simple benzoate ion tends to be unfavorable (makes ΔF_{rel} more positive by several hundred calories), if the slightly bulky chlorine atom is not able to go into the position intended for the azo group of the original antigen. When the much larger *p*-hydroxyphenylazo group is present in the hapten, it occupies the position intended for the protein of the original antigen; the introduction of a chlorine atom into such a hapten then invariably decreases its affinity for the antibody site. The amount of the decrease in affinity is different, depending upon the location of the chlorine atom on the benzene ring. This reveals something about the nature of different portions of the surface of the hapten-binding site on the antibody. It is clear that experiments of this kind allow one to probe the various regions of the antibody-combining sites and to obtain quite detailed information on the factors that determine antibody-antigen specificity.

2. Enzyme-Substrate Interactions

Enzymes are well known to be rather particular about the substrates whose reactions they choose to catalyze. The degree of this specificity varies widely, however. The enzyme urease evidently catalyzes only the hydrolysis of urea and appears to have no noticeable effect on any urea derivative or other substance that has yet been tested. Fumarase adds water to the double bond of the fumarate ion to produce L-malate ion and no other double-bonded organic molecule has been found which is attacked by this enzyme. On the other hand, certain lipases will hydrolyze ester bonds containing widely different organic groups attached to either side of the bond.

It is generally agreed that enzymes act by adsorbing the substrate molecule or molecules at some "active site" on the enzyme molecule in such a way that the appropriate bonds in the substrate show an increased reactivity. It is difficult to avoid thinking in terms of the hypothesis that the enzyme-substrate interaction, like the antibody-antigen interaction, depends upon a complementarity in structure between the two substances, the substrate molecule fitting into a characteristically shaped impression in the surface of the enzyme molecule. The increased reactivity may arise because in the adsorbed state the bonds involved in the

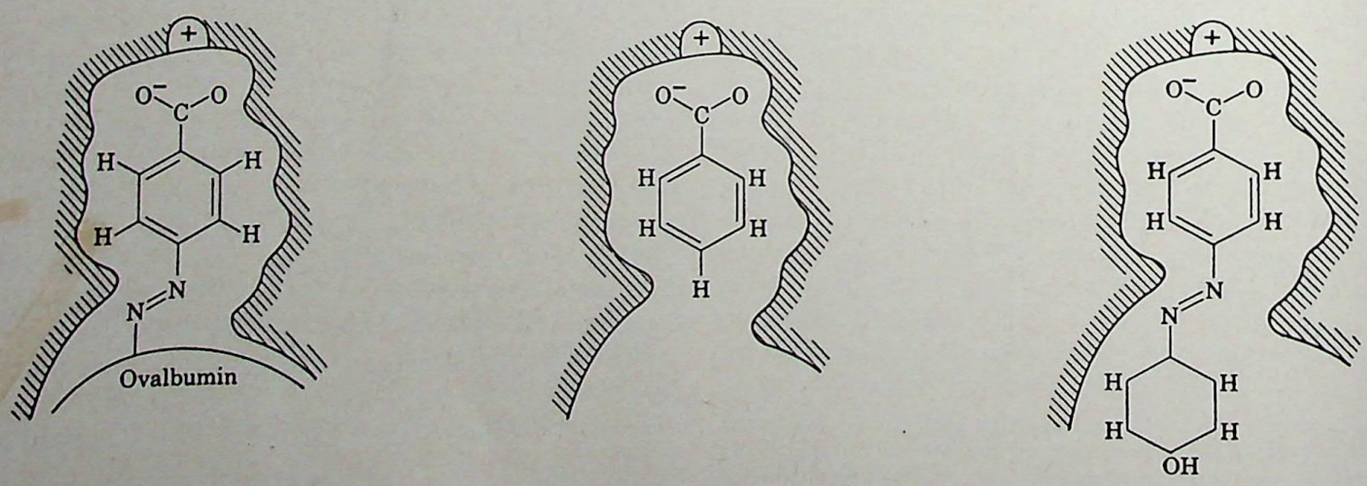


FIG. 3. Hapten inhibition of anti-X_p-globulin.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

reaction are placed under strain or because they are brought into a favorable position for reaction. This simple picture seems to be consistent in a general way with the wide range of specificity found among enzymes. Highly specific enzymes presumably contain a highly characteristic impression of the substrate molecule into which it is not possible to fit even closely related molecules capable of undergoing the same reaction as the substrate. Enzymes showing a wide range of specificity, on the other hand, must interact with only the region of the substrates in the immediate vicinity of the bonds that are involved in the reaction. This picture also explains why substances closely related to a given substrate, but incapable of undergoing the reaction that is catalyzed by the enzyme, often act as powerful inhibitors (e.g., malonic acid and malic acid strongly inhibit the conversion of succinic acid to fumaric acid by succinic dehydrogenase). These inhibitors must fit into the site intended for the substrate but, like the dog in the manger, they neither take advantage of the catalytic capabilities of the site nor permit access to the true substrate.

The hypothesis of structural complementarity also explains the high degree of stereospecificity shown by practically all enzyme reactions in which the potential substrates are capable of existing as stereoisomers. (E.g., the dehydration of malic acid in the presence of fumarase to form fumaric acid is absolutely specific for L-malic acid, D-malic acid being unaffected by fumarase to any detectable degree; similarly, maleic acid, the *cis* isomer of fumaric acid, is not hydrated by fumarase to any detectable degree. Most peptidases will not hydrolyze peptide bonds involving D-amino acids.) Suppose that the active center of an enzyme is surrounded by a mold into which all of the chemical groups in the vicinity of a double bond of the *trans* isomer of a compound will fit. It is easy to understand why this mold will not be able to accommodate the *cis* isomer of the compound, which has an entirely different external shape. The specificity to optical isomers is no less easy to understand in terms reflecting the common experience that a left hand will not fit into a glove made to cover its mirror image, the right hand.

It is interesting to examine some typical enzyme systems in order to observe the types of behavior that can be accommodated to the complementarity concept. One finds that many systems can be interpreted very easily in these terms, but one finds also that there are certain instances in which the concept does not present a very convincing picture of what must be going on.

(*a*) *Peptidases*

Bergmann studied the action of various peptidases on simple peptides and was able to show that there is a high degree of specificity among these enzymes which depends on the nature of the groups located on either side of the peptide bond that is split. He was able to formulate certain general specificity requirements for the action of pepsin, trypsin, and chymotrypsin, which are summarized in Table III. Further work has shown, however, that many of these requirements are not met by all substrates of these enzymes, and the specificity picture has become somewhat more complex (see, for example, the discussion in Dixon and Webb[6] and Neurath and Schwert[7]). To illustrate these complexities, the chief points of attack on ribonuclease by various peptidases are shown in Fig. 4. Inspection of this figure and Table III will show that the specificity discovered by Bergmann for trypsin is invariably maintained: the points of attack by trypsin are always at the peptide bond next to either arginine or lysine, with the carbonyl end of the bond that is hydrolyzed belonging to the cationic amino acid. It does appear, however, that the bond adjacent to one of the cationic amino-acid residues of ribonuclease (the lysine at position 44 from the lysine end of the chain) may not be attacked by trypsin. The points of attack on ribonuclease by chymotrypsin are frequently adjacent to aromatic side chains (usually at the carbonyl side), as demanded by the Bergmann rules. Unexpected exceptions are found, however, in the ready opening of the leucine-threonine linkage between positions 35 and 36, and of the alanine-lysine linkage between positions 103 and 104. A weak ability of chymotrypsin to split peptide bonds involving methionine had been noted in simple peptides (chymotrypsin also hydrolyzes a cystine-serine bond in insulin). These exceptions are somewhat puzzling from the point of view of the complementarity concept, because the groups that are involved can hardly be said to resemble in general shape the aromatic groups that the Bergmann rule demands for chymotrypsin. More complex factors seem to operate than the mere binding by the enzyme of the amino-acid residue next to the peptide bond that is being hydrolyzed. The slight ability of methionine and cystine to replace an aromatic ring in some substrates for chymotrypsin is especially puzzling and might suggest that an electronic mode of interaction may be involved rather than one involving steric complementarity.

TABLE III. Bergmann's specificity rules for the hydrolysis of the peptide bond in $R_2—NH—CHR_1—CO—NH—R_3$.

Enzyme	R_1	R_2	R_3
Pepsin	glutamic or aspartic	not H	aromatic (tyrosine or phenylalanine)
Trypsin	cationic (lysine or arginine)	not H	nonspecific; may be H
Chymotrypsin	aromatic (tyrosine or phenylalanine)	nonspecific	nonspecific

The specificity rule for pepsin proposed by Bergmann on the basis of its action on simple peptides turns out to be violated frequently when pepsin acts on long

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

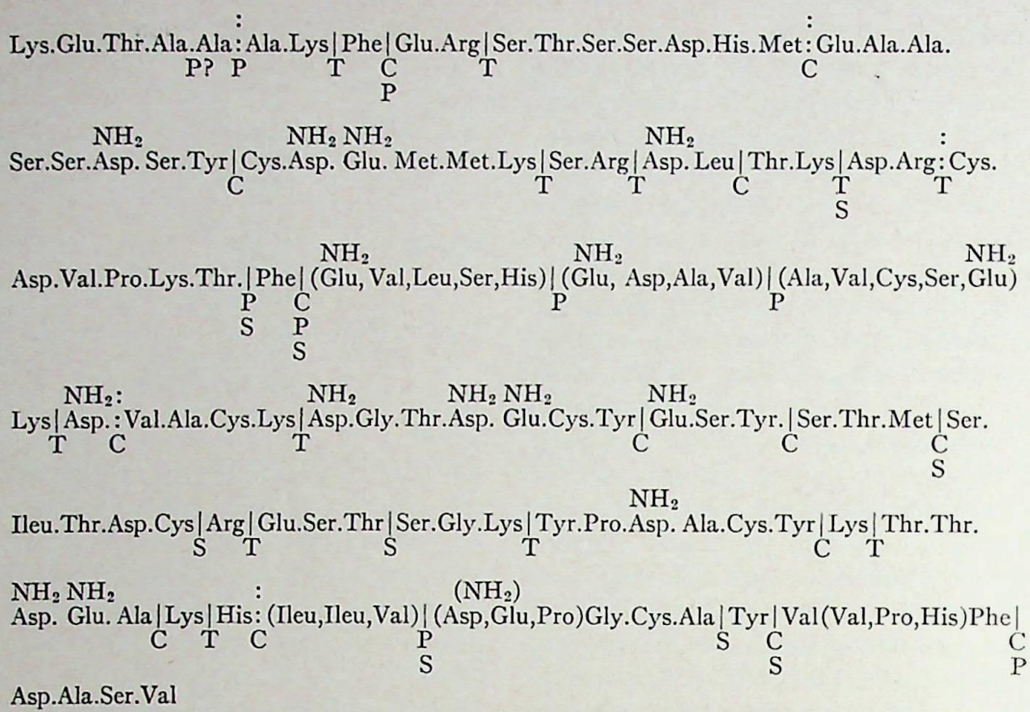


FIG. 4. Points of hydrolysis of ribonuclease by chymotrypsin (C), pepsin (P), trypsin (T), and subtilisin (S). Dotted lines represent points of slower hydrolysis.

polypeptide chains. This is evident from Fig. 4, where the attack on an alanine-alanine bond may be noted. Pepsin also attacks an alanine-serine bond in ACTH and a leucine-valine linkage in insulin. In general, however, pepsin does show a tendency to attack either bonds involving glutamic or aspartic acids by themselves, or bonds involving aromatic amino acids by themselves—in partial agreement with Bergmann's rules. It appears that the true range of specificity of pepsin does not reveal itself until one allows it to act on relatively long polypeptides.

It is interesting that trypsin and chymotrypsin are able to hydrolyze esters and suitably activated carbon-carbon bonds as well as peptides. The specificity requirements on the amino acids adjacent to the ester or carbon-carbon bond are similar to those observed for peptides with the same two enzymes. Here one has a particularly striking example of the importance of the environment of a reacting bond, rather than the bond itself, in manifesting the specificity of the enzyme.

(b) *Cholinesterase and Acetylcholinesterase*

The reaction

$$(CH_3)_3N^+-CH_2-CH_2-O-CO-CH_3 \quad \text{(acetylcholine)}$$
$$+H_2O \rightarrow (CH_3)_3N^+-CH_2-CH_2-OH \quad \text{(choline)}$$
$$+CH_3-CO-OH$$

is catalyzed by two distinct groups of enzymes which have interestingly different specificity requirements. One of these groups of enzymes, the acetylcholinesterases is inhibited by the substrate acetylcholine at higher substrate concentrations, so that its activity *vs* substrate-concentration curves pass through a maximum. The other group of enzymes, known as cholinesterases, is not so inhibited, and the activity *vs* substrate-concentration curves follow the normal Michaelis-Menton behavior. Both kinds of enzymes can act on many substrates related (in some instances rather distantly) to acetylcholine, but mention is made only of the interesting and opposite effects of replacing the acetyl group with other acyl groups containing longer hydrocarbon chains. In mammalian brain acetylcholinesterase, the replacement of acetyl with butyryl greatly lowers the activity and does not remove the inhibition at high substrate concentrations. In cholinesterase, on the other hand, the same substitution greatly increases the activity of the enzyme. Here one has an example of two enzymes which catalyze the same reaction but whose specificity patterns are markedly different, showing that considerably different types of complementarities must be possible for a given substrate.

(c) *Glycosidases*

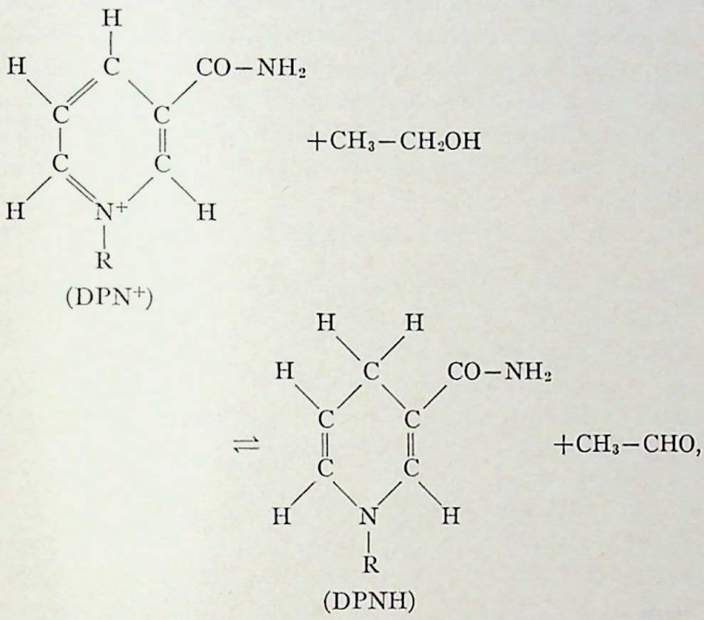
The hydrolysis of α- or β-alkyl- or aryl-glycosides to form free sugars plus alcohols or phenols is catalyzed by a group of enzymes that show interesting specificities. In general, a given glycosidase is highly specific to the sugar moeity, and to whether the glycoside has the α- or β-configuration, but it is relatively unconcerned about the nature of the aglucone residue. For instance, α-glucosidase will hydrolyze α-methylglucoside, α-phenylglucoside, and many other α-glucosides, but it will not cause the hydrolysis of any β-glucoside or any α-glycoside of a sugar other than glucose. Here one has

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

examples of very high specificity toward a group that lies on one side of the chemical bond that is attacked, and a very low specificity to the other side.

(*d*) *Stereospecificity in the Transfer of Hydrogen to DPN in Dehydrogenase Reactions.*

Vennesland[8] and Westheimer have shown by means of ingenious experiments with deuterium that, in the reaction



both the addition of the hydrogen atom to the pyridine ring of DPN^+ in the forward direction and the addition of hydrogen to acetaldehyde in the reverse reaction are stereospecific. Thus, if the equilibrium

$$CH_3-CH_2-OD+DPN^+ \rightleftharpoons CH_3-CHO+DPNH+D^+$$

is carried out in D_2O, it is found that no additional deuterium atoms are introduced into either DPN^+, DPNH, alcohol, or aldehyde. Similarly, no normal hydrogen is introduced when the reaction

$$CH_3-CD_2-OH+DPN^+ \rightleftharpoons CH_3-CDO+DPND+H^+$$

is equilibrated in H_2O. This proves that the hydrogen atom is transferred directly from the carbon atom of the alcohol to the DPN^+ (and vice versa) and does not go through the solvent. Furthermore, the hydrogen atom introduced into the DPNH must lie on one definite side of the pyridine ring. The ring must lie, therefore, in a definite orientation on the enzyme surface, and relative to the alcohol molecule, in the complex of the enzyme with its two substrates. That the orientations of the ethanol and aldehyde molecules also must be fixed rigidly relative to the pyridine ring is shown by the fact that the reaction

$$H^+ + DPND + CH_3-CHO \rightarrow DPN^+ + CH_3-CHD-OH$$

produces a single optical isomer of ethanol-1*d*. Here is an especially clear example of the detailed stereochemical specification of the structure of an enzyme-substrate complex. It must take advantage of a large fraction of the possible complementarity that can exist between the substances concerned in the reaction.

(*e*) *The Koshland-Stein Rule*

Koshland and Stein[9] have proposed the general principle that "if the enzyme requirements for a substrate having an oxygen bridge, i.e., R—O—Q, show high specificity for R and low specificity for Q, than R—O cleavage occurs in the enzyme reaction." This rule has been found to be valid for invertase, alkaline phosphatase, acid phosphatase, trypsin, and chymotrypsin. The simplest interpretation of the rule in terms of complementarity would be that the group with the highest specificity would be presumed to have the closest contact with the enzyme; the enzyme then would be expected to have the greatest activating effect on the bond adjacent to this group.

(*f*) *Existence of Identical Amino-Acid Sequences in the Active Centers of Enzymes Having Diverse Specificities*

It has been found recently that the active centers of thrombin, chymotrypsin, trypsin, and phosphoglucomutase all contain the sequence aspartic acid-serine-glycine. These enzymes have markedly different specificities, the first three catalyzing the hydrolysis of peptide bonds, whereas phosphoglucomutase catalyzes the transfer of a phosphate group from one glucose ring to another. The specificity of these enzymes must clearly reside in regions quite distinct from the center that activates the chemical bonds which are disrupted.

3. Binding of Small Molecules to Proteins (Other Than Hapten-Antibody Interactions)

Serum albumin is remarkable for its ability to adsorb a wide variety of small molecules, particularly (though not at all exclusively) those bearing negative charges (acetate ion, halide ions, methyl orange, and many other anionic dyes, detergents, etc.). The most striking feature of this adsorption, however, is the lack of any reasonably marked specificity. Changes in the structures of the adsorbed molecules do have some effects on the strength of the binding. These changes in affinity are not, however, as large as one would expect to find if the surface of the serum albumin molecule were covered with a mosaic of rigid binding sites having a reasonably high degree of complementarity to the necessarily limited number of types of substances which such a mosaic might potentially bind. (For example, enantiomorphic isomers of a dye are adsorbed to almost the same extent.) In order to account for this, Karush[10] has proposed the concept of "configurational adaptability," according to which the serum-albumin molecule is considered to possess a degree of internal flexibility.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

As a result, the surface conformation can be changed so as to establish a measure of complementarity with the surface of almost any molecule that approaches it, and there is no need to deal with a limited number of fixed types of potential binding sites. A strong point in support of this concept is the fact that serum albumin does possess a high degree of internal flexibility as compared with other proteins, which also show a far smaller affinity for small molecules.[11] In a sense, the serum-albumin molecule behaves like a lump of putty onto which one can stick an unlimited number of different shapes.

The concept of configurational adaptability is important because it shows that complementarity need not imply specificity. Furthermore, it raises the possibility that, in some instances, the specific complimentarity structure may be developed only when the substrate is present.

SPECIFICITY AND ORGANIZATION

Green has set forth[12] reasons for believing that, in certain biochemical processes involving a sequence of steps, each catalyzed by a different enzyme, the enzyme molecules must occupy fixed positions in space which are located so as to facilitate the successive reactions. He speaks of "an organized mosaic of enzymes in which each of the large number of component enzymes was uniquely located to permit efficient implementation of consecutive reaction sequences." A primary example of this mosaic is the cyclophorase system, which includes all of the enzymes for the citric-acid cycle, fatty-acid oxidation, oxidative phosphorylation, and terminal electron transport. This system of enzymes has been shown to be located exclusively in the mitochondria. In view of the striking mitochondrial structures which have been revealed by electron microscopy [see the papers by Sjöstrand (p. 301) and Fernández-Morán (p. 319) in this volume], there is a certain temptation to accept this theory.

Dixon and Webb[13] have disputed the necessity for assuming organization in this spatial sense. As an alternative, they go back to the point of view presented twenty-five years ago by Hopkins[14] in these words: "The organizing potentialities inherent in highly specific catalysts have not, I believe, been adequately appraised in chemical thought. Highly specific catalysts determine just what particular materials, rather than any others, shall undergo reaction. They select from their environment. The specific catalyst determines which among possible paths the course of change shall follow. It has directive powers . . . ." Dixon[15] speaks of "organization by specificity" as an alternative to Green's spatial organization of enzyme systems. The concentration of the cyclophorase enzymes into the mitochondria is ascribed to the need to reduce the distance between the enzymes involved in consecutive reactions. The cell is able in this way to reduce the "transit times" required for the products of one enzymatic step to diffuse to the enzyme molecules involved in the next step. According to this view, there is no need to assume any regular arrangement of the different enzymes. It is necessary only to have an arrangement (a random one would do) with a sufficiently small mean distance of separation between the enzymes involved in the different enzymatic steps.

It is interesting in this connection to estimate the order of magnitude of the transit time for a typical metabolite to move a given distance from one enzyme to another by a diffusion mechanism. It is well known that if x^2 is the mean square distance moved in time t by a particle whose macroscopic diffusion constant is D, then

$$x^2 = 2Dt.$$

Typical metabolites have diffusion constants of the order of 1 cm^2/day, or 10^{-5} cm^2/sec. Thus, it is found that, if the enzymes required for two successive steps of a reaction sequence are 1000 A apart, the transit time is 10 μsec. If the separation is 100 A, the transit time is 0.1 μsec. It would appear, therefore, that the arrangement of enzymes of a reaction sequence in layers whose repeat distance is a few hundred Ångström units, without any regard to the serial positions in space of the enzymes involved in successive reaction steps, would give a more than adequately short transit time for most biochemical requirements. Specificity alone should be adequate to cope with problems of organization at this level. This is not to say, of course, that the living cell actually is organized in this manner. For reasons unknown, it may have chosen to solve its transit time problems in a less straightforward manner.

BIBLIOGRAPHY

[1] K. Landsteiner, *The Specificity of Serological Reactions* (Harvard University Press, Cambridge, 1945), revised edition.

[2] W. C. Boyd, *Fundamentals of Immunology* (Interscience Publishers, Inc., New York, 1956), third edition.

[3] K. Landsteiner, and H. Lampl, Biochem. Z. **86**, 343 (1918)

[4] D. Pressman, M. Siegel, and L. A. R. Hall, J. Am. Chem. Soc. **76**, 6336 (1954).

[5] L. Pauling, D. Pressman, and A. L. Grossberg, J. Am. Chem. Soc. **66**, 784 (1944).

[6] M. Dixon and E. C. Webb, *Enzymes* (Academic Press, Inc., New York, 1958), p. 251.

[7] H. Neurath and G. W. Schwert, Chem. Revs. **46**, 69 (1950).

[8] B. Vennesland, J. Cellular Comp. Physiol. **47** (suppl. I), 201 (1956).

[9] D. E. Koshland and S. S. Stein, J. Biol. Chem. **208**, 139 (1954).

[10] F. Karush, J. Am. Chem. Soc. **72**, 2705 (1950).

[11] W. Kauzmann in *Symposium on the Mechanism of Enzyme Action*, W. D. McElroy and B. Glass, editors (The Johns Hopkins Press, Baltimore, Maryland, 1954), p. 70.

[12] D. E. Green, Symposia Soc. Exptl. Biol. **10**, 30 (1957).

[13] See reference 6, p. 242.

[14] F. G. Hopkins, Proc. Roy. Soc. (London) **B112**, 159 (1932).

[15] M. Dixon, *Multienzyme Systems* (Cambridge University Press, London, 1949).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

REVIEWS OF MODERN PHYSICS VOLUME 31, NUMBER 2 APRIL, 1959

60

Blood Coagulation—A Study in Homeostasis

DAVID F. WAUGH

Department of Biology, Massachusetts Institute of Technology, Cambridge 39, Massachusetts

IN this discussion, a brief account is to be made of one of the mechanisms involved in homeostasis, a word which means that the chemical and physical processes occurring within the fluid matrix or internal environment of an organism are integrated so that a stable physiological level of each is maintained. Thus, if one process deviates from its stable level, other mechanisms are shifted slightly or considerably, as the occasion demands, in an attempt to bring this process and the entire system back to its normal state. Of course, everyone appreciates the fact that the closely related set of activities occurring in one physiological system are influenced by the sets of activities occurring in other systems. For example, the system discussed herein, the blood coagulation system, is influenced markedly by stress (as, of course, are the other physiological systems.) The mechanisms in one physiological system, then, are not at liberty to establish levels of activity which are optimal for that system as an isolated system: the levels are a compromise derived from the activities of the organism as a whole.

It should be made clear at the start that the information presented here is also the result of a compromise: it represents the activities in a chosen individual only in a general way and is a simplification most parts of which can be challenged. This situation stems from the fact that, while there is general agreement about many of the coagulation mechanisms, others have been recognized as possibilities only recently and an understanding of details is incomplete. Extensive use has been made of certain references in preparing this discussion: the books by Wintrobe[1] and Biggs and MacFarlane,[2] the review articles by Brinkhous, Langdell, and Wagner,[3] and Seegers,[4] and the symposium edited by Brinkhous.[5]

Applied to the blood vascular system the word homeostasis would include first, the hemostatic mechanisms—that is, the mechanisms which are involved in sealing a wound which has opened the usually closed circulatory system—and second, the more mysterious set of mechanisms which are involved in maintaining the thin fragile walls of the capillaries in such a state that the proteins and formed elements (red and white cells) of the blood do not escape from the circulation and pass into the tissue spaces. The first set of mechanisms, the hemostatic, involve the closing of small vessels through vascular constriction, plugs of platelets, and coagulation in which the fluid plasma is transformed into a gel. We can be concerned only with coagulation, where it is found that an extensive set of interlocking processes exist in what is quite evidently a poised equilibrium such that a displacement of the equilibrium in the proper direction leads to a rapid autocatalytic development of the clotting enzyme thrombin from its inactive precursor prothrombin. A general scheme of the features of the coagulation system to be discussed here is shown in Fig. 1. The explosive development of thrombin swamps out, locally, all of the anticlotting mechanisms and a clot develops, the latter being the result of a change in the structure and in the state of aggregation of the protein fibrinogen. The clot is confined locally by a set of slower reactions, some of which are autocatalytic and controlled by thrombin development and most of which are backed up by large capacities. These reactions nullify the clot-developing system at several points.

The trigger for the mechanisms which displace the poised equilibrium is apparently disruption of a boundary which isolates key substances, either within small blood elements, the platelets, or within the tissue cells. On wounding, the intermixing of blood and tissue juices and the presence of rough surfaces at the site of injury, which cause platelet decomposition, are sufficient to give the rapid development of thrombin described in the following.

One might venture that if hemostasis were the only objective of the coagulation system it could be much simpler than it is. Many of the extraordinary complications are probably associated with the fact that coagulation must occur continuously at low levels in order to preserve the normal structural and permeability characteristics of the capillaries. The physiological necessity for slow continuous clotting is referred to later in an attempt to justify this statement. It is just mentioned now, however, to focus attention at the start on the possibility that one is dealing with a system which, on the one hand, must operate at a low but precisely controlled level of activity but which, on the

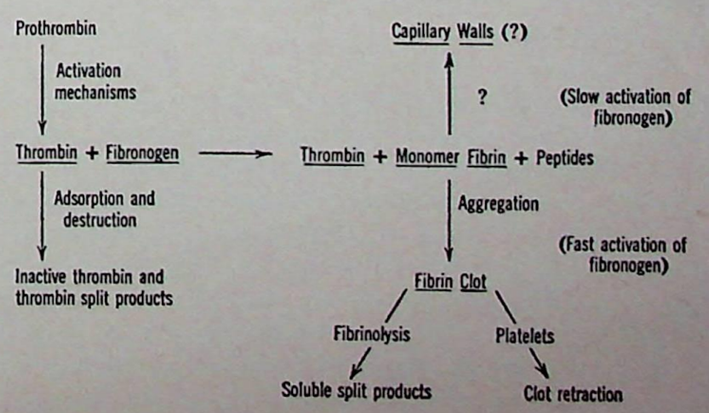


FIG. 1. General scheme.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

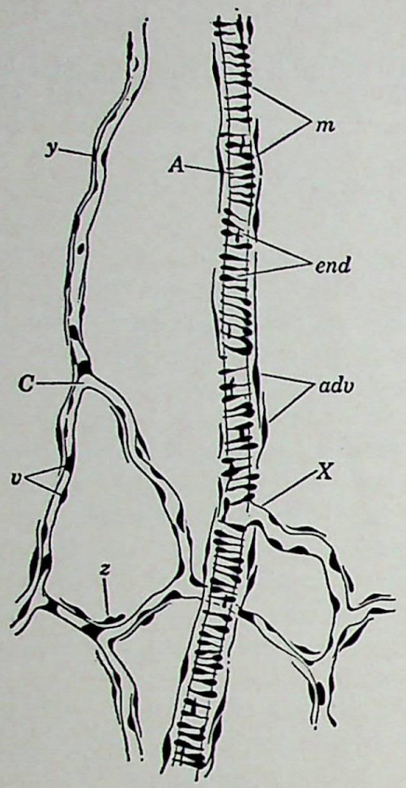


FIG. 2. Small artery, *A*, and capillaries, *C*, from the mesentary of a rabbit; *M*, muscle cells in the media; *adv*, adventitia; *X*, origin of a capillary from the artery; *y*, pericyte; *z*, perivascular histiocyte; *end*, *v*, endothelial nuclei [from A. Maximow and W. Bloom, *A Textbook of Histology* (Saunders Company, Philadelphia, 1942), fourth edition].

other hand, can be triggered into explosive activity which is not controlled but must be confined.

Blood flows within a set of tubes of varying size, the larger tubes (arteries, arterioles, veins, venules) being covered by substantial layers of muscle and connective tissue. The smallest vessels, the capillaries, are our chief concern. These fragile vessels (Fig. 2) have a wall which is constructed of a single layer of flat endothelial cells cemented at their boundaries and invested by a thin network of connective-tissue fibers. The caliber of the capillaries is close to the diameter of a single red cell, and is thus about 8 μ in man. It is believed that water and solutes generally pass through the cytoplasm of the endothelial cells. White cells may migrate through the substance of the endothelial cells or through the boundaries between cells.

Blood[1] itself is a fluid 40% of whose volume is made up of formed elements (red cells, white cells, etc.) and 60% of a straw-colored fluid plasma containing seemingly an endless variety of compounds, many in the process of being transported from one organ to another. The most important of the formed elements in this discussion is the platelet, a disk-shaped body about 2 μ in diameter. The platelets can become sticky and adhere to each other forming a mass which itself is of importance in stopping the flow of blood from capillaries. In addition the platelets, as mentioned, normally act as unavailable depots for important coagulation components. The platelet is fragile. If it comes in contact with a surface foreign to the circulation (i.e., glass) it adheres, spreads out, and bursts liberating its contents.

Within the plasma are found a series of proteins and lipoproteins all involved directly or indirectly in clotting. These proteins are described as they come up for discussion, the current series includes prothrombin, thrombin, and fibrinogen. Prothrombin, in the purified preparation of Seegers *et al.*,[6] appears as a molecule of $M \sim 66\,000$ and as a prolate ellipsoid may be described as having a length of 119 A and a diameter of 34 A.[7] This molecule can dissociate on dilution into halves. It is quite stable in the pure state and stable in plasma when prothrombin activators are absent. Prothrombin, as its name implies, has no action whatsoever on fibrinogen.

Thrombin is one of the derivatives of prothrombin. It can be obtained from prothrombin by a variety of techniques, the important one in this instance being the normal biological-activation process involving a variety of substances. In this activation process, the molecular weight of prothrombin does not appear to change radically. In the concentrated solutions required for physical studies, however, thrombin may be associated. Our current studies suggest that the minimum molecular weight of thrombin is about 22 000. One unit of thrombin is defined as that amount which clots plasma or a solution of purified fibrinogen corresponding to plasma in 16 sec. Since preparations of thrombin containing 2000 units/mg have been made, it is clear that a single unit is represented by about 5×10^{-4} mg. There is sufficient prothrombin in one milliliter of plasma to give 300 units of thrombin, corresponding to the small amount of about 0.15 mg of prothrombin per milliliter of plasma. The total possible number of thrombin units is, however, staggering. Immediate activation of a few percent of the prothrombin would coagulate blood in seconds. In the coagulation process, thrombin acts as an enzyme which splits an acidic peptide from fibrinogen (Fig. 1).

Fibrinogen has a molecular weight of 340 000 and is an asymmetric protein of axial ratio about 10.[8] It has an interesting substructure[9] which cannot be discussed here for lack of space. In pure form, it is a fairly stable protein which survives rough handling. After thrombin has split off an acidic peptide, the reactivity of the original protein is altered, the fibrin monomer produced by this enzymatic alteration having now a marked capacity to interact in a relatively specific fashion with other fibrin monomers. At the same time, a small amount of stabilizing plasma globulin is incorporated.[10] The interaction leads predominantly to end-wise association which produces fibrin fibrils or fibers having asymmetries orders of magnitude greater than that of the original fibrinogen molecule. During growth of the set of fibrin fibrils there occurs a brief interval of time during which the total volume changes from being greater to being less than the sum of the volumes dominated by each fibril. The

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

fibrils start to overlap and to crosslink where they come into contact. In this way a space filling network of strands develops.[11] At first, the strand network is tenuous and easily ruptured, but as old strands become thicker and as new strands are incorporated into the network, the clot progressively develops rigidity. As is seen later, this is not the end of the story, for, in the presence of platelets, a spontaneous squeezing out of fluid may occur (clot retraction). This is an important activity of the clot in forming a compact closure over a wound. If retraction and drying of the clot do not occur, an enzyme plasmin may be activated from its precursor.[12] Apparently, the enzyme plasmin has been devised specifically to dissolve fibrin.

The first portion of this general scheme to consider is the transformation of prothrombin into thrombin. This phase of the over-all problem has received a great deal of attention, the natural result of the fact that discoveries in clinics have revealed an ever-increasing series of patients whose plasmas have lacked one or another of the activators necessary for the prothrombin-thrombin conversion, and of the fact that prothrombin has been prepared in relatively pure form, first by Seegers and more recently by others, and its activation has been studied under a variety of conditions.

The scheme which is shown in Fig. 3 is a simplified version of experience in the field.[1–5] The central theme is the conversion of prothrombin into thrombin. The top half of the figure depicts clotting events which can occur when blood is drawn and every effort is made to keep the blood free of contamination with tissue components. Before proceeding, attention is directed to the central role assigned here to thrombin not only as the clotting enzyme but also as an agent which feeds back and promotes the formation of more of itself. As can be seen, this feedback enters at several levels into the set of interactions and, in order to promote the explosive

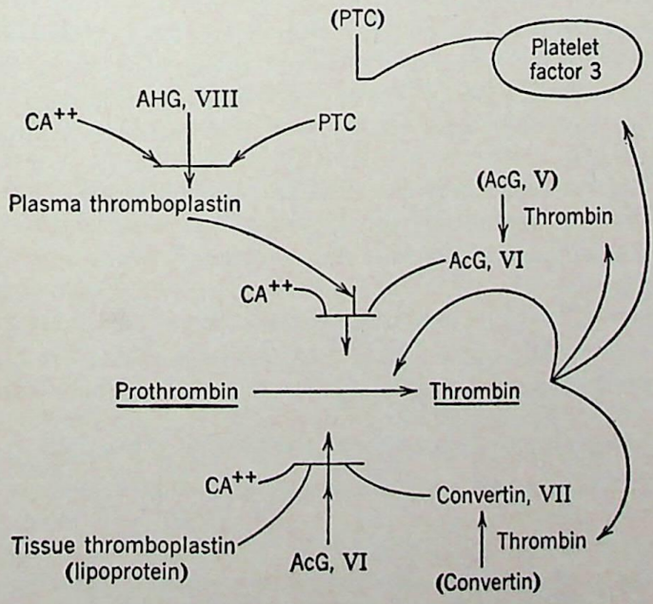


FIG. 3. Illustrative scheme for mechanisms involved in prothrombin activation.

liberation of thrombin, what is necessary, so far as this portion is concerned, is platelet decomposition and a nudge in the direction of thrombin production.

To start with, on being drawn, the blood comes into contact with a foreign surface and certain of the platelets respond by rupturing and liberating a series of compounds, the one singled out being platelet factor 3, a lipoprotein. This factor combines with or activates an inactive form of a substance which is probably a protein called plasma thromboplastin component (PTC). The inactive forms of components have been placed in parentheses, the Roman numerals indicate the accepted number in the "factor" series. Plasma thromboplastin component now interacts with antihemophilic globulin (AHG) and calcium ion to give plasma thromboplastin. This is not sufficient in itself to produce rapid prothrombin conversion, at least one additional step being necessary—namely, an interaction with the active form of accelerator globulin (AcG, VI). A rapid conversion of prothrombin now takes place. Note how thrombin is instrumental in converting inactive accelerator globulin to an active form, that it is involved in platelet alterations and may be involved directly in the conversion of prothrombin. All of the elements in this scheme are present in plasma.

The lower part of the Fig. 3 shows a sequence of events involving tissue components (tissue thromboplastin). Tissue thromboplastin short circuits platelet contributions at the start but requires calcium ion, the active form of accelerator globulin, and a new protein factor, convertin, which is produced from a precursor by the action of thrombin (cf. Goldstein and Alexander in reference 5, p. 93). It is generally held that the tissue thromboplastin-convertin cycle gives an immediate thrombin production which then operates on itself and on the upper mechanisms of Fig. 3 to yield a fast production of thrombin.

During the rapid development of thrombin (upper part of Fig. 3), all of the prothrombin may not be converted to thrombin. In the scheme presented, this is most readily understood by the fact that, during prothrombin activation, plasma thromboplastin, and thus at least some of the components which give rise to plasma thromboplastin, are either inactivated or consumed, as is true also of tissue factors. This is related to the confinement mechanisms which are outlined in the following.

Another aspect of the thrombin-production feedback mechanism which should be mentioned is that the platelets are thought to liberate a substance which is an antiheparin, and since, as is seen, heparin is intimately involved in the destruction of thrombin and thromboplastin, removal of heparin constitutes an additional positive factor in thrombin production.

The scheme of Fig. 3 has presented one possible sequence of events. The situation is probably more complicated than this, as is evident from a consideration of the following. First, at least two possible factors have

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

been omitted from the scheme. These are the Hegeman factor and PTA or plasma thromboplastin antecedent. Second, the number and sequences of reactions leading to the development of plasma thromboplastin are not clear. Not only is it possible that the details of factor interactions as given are incorrect, but also there is known to be opportunity for substitution under the proper conditions. For example, the platelets liberate a material which can substitute for the active form of accelerator globulin.

It is in this series of reactions that one finds the proven genetically controlled clotting defects. An abnormality in antihemophilic globulin is the defect of classical hemophilia, which was described over 150 years ago. It is a sex-linked recessive and manifests itself in varying deficiency. Seegers *et al.*[13] have evidence to show that this abnormality may be owing to inhibitors which block AHG. Lack of accelerator globulin (V), although it may be acquired through liver damage, may also be of hereditary origin; however, the inheritance is uncertain. Lack of PTC (hemophilia B, Christmas disease, etc.) is an hereditary disease which is a sex-linked recessive, similar to AHG deficiency.

There are two important series of reactions which are involved in confining coagulation to a locality and in removing unwanted clots. In approaching the mechanisms which confine clot development, it is necessary to realize that in wounds the hemostatic mechanisms are operating in the relatively stagnant or slowly flowing blood which occurs near the constricted ends of cut vessels. Whatever blood pressure is operating on these cut vessels will tend to work the coagulating system to the exterior. Thus, the problem is one of handling back-diffusion and any back-flow of active clotting agents. Of course, if such agents are carried into the general circulation, there occurs a relatively extensive dilution.

Confinement of clot development and the shutting off of coagulation reactions involve a series of antieffects which operate on the basis of adsorption, stoichiometric combination, and enzyme action. Some mechanisms are not well established, although a high probability can be assigned to their existence.

The first rather interesting reaction which is known to have importance in limiting clot development is the adsorption of thrombin to fibrinogen and fibrin.[3,4,12,13] In a static blood volume, such an effect allows complete local clotting to be accomplished before thrombin can diffuse out of the locality.

There then follows a series of reactions which have been grouped in Table I according to whether or not they are thrombin dependent. The second group is thrombin dependent; the equations mean that thrombin, alone or with other materials, will destroy three key procoagulants: accelerator globulin, prothrombin, and antihemophilic globulin.

What is implied by the three equations listed in the third group is that for every coagulant in the active form there is a factor more or less specifically designed to effect its inactivation. When the mechanisms of the first three groups of Table I are coupled with the many observations that materials such as AHG, PTC, AcG, platelet factor 3, etc., appear to be *consumed* in activating prothrombin, it is apparent that the explosive development of thrombin can be sharply terminated in the region of the clot.

The presence of sizable amounts of free thrombin always, however, constitutes a dangerous situation and two mechanisms, one rapid and the other slow, have apparently been devised to cope with the situation. The first of these is indicated in the first equation under group 4 and is the rapid reaction. The experiments designed to prove the existence of this inactivating factor are assailable, but essentially the suggestion is that, when normal prothrombin is activated to thrombin in plasma, there arises at the same time an ephemeral factor which can inactivate the resulting thrombin.

The greatest capacity for inactivating thrombin is distributed throughout the plasma itself. This activity is indicated in the second reaction under group 4, the responsible substance or substances being referred to as normal antithrombin. There is present in plasma sufficient normal antithrombin to inactivate 1.5 times the thrombin which could appear if all of the prothrombin were activated to thrombin. This places the capacity of antithrombin at about 450 units/ml. Under ordinary conditions, normal antithrombin acts relatively slowly, as is evident from the time disappearance of thrombin added to plasma which has been diluted fifty-fold with saline (Fig. 4, curve 1, see Waugh and Fitzgerald[15]). It is certain that antithrombin is itself inactivated in the process of destroying thrombin; the reaction is thus stoichiometric. A reversal of the thrombin-antithrombin complex has so far *not* been accomplished. It should be emphasized once more that the capacity of plasma antithrombin is about 450 units/ml. This large plasma-antithrombin capacity to a certain extent compensates for the small rate constant of the second-order thrombin-antithrombin reaction, the result being that when one unit of thrombin is present per milliliter plasma, inactivation will take place at the rate of 0.05 units

TABLE I. Confinement and shut-off mechanisms.

Group 1.	Thrombin + fibrin $\rightleftarrows$ thrombin fibrin
Group 2.	AcG, VI $\xrightarrow{T}$ inactive AcG, VI
	Prothrombin $\xrightarrow{T, \text{factors}}$ inactive prothrombin
	AHG $\xrightarrow{T}$ inactive AHG
Group 3.	Antithromboplastin + thromboplastin → inactive product
	Antiplatelet factors + platelet factors → inactive products
	Anti AcG VI + AcG VI → inactive product
Group 4.	Thrombin + inactivating factor → inactive thrombin
	Thrombin + antithrombin → inactive (thrombin antithrombin)

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

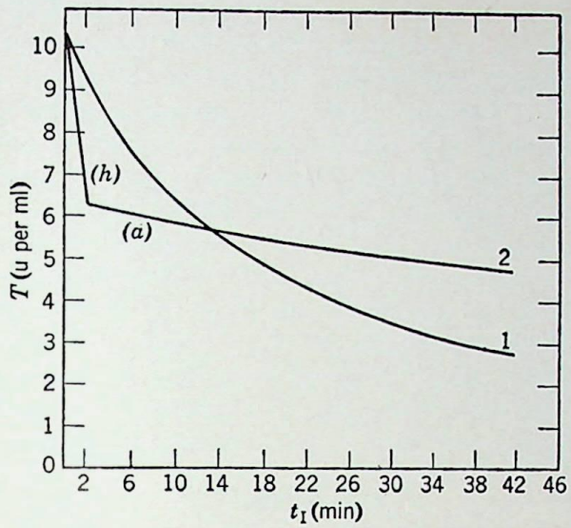


FIG. 4. Inactivation of thrombin by antithrombin. Curve 1 refers to normal antithrombin and curve 2 to the effects which are produced by traces of heparin.

thrombin per second. This rate of inactivation will increase the clotting time under conditions of fast activation but will not act with sufficient rapidity to prevent clotting.

The use of heparin, the highly acidic (sulfated) mucopolysaccharide, as an anticoagulant is well known. At this time, attention is directed only to the fact that heparin, at the small concentrations at which it exerts a strong anticoagulant action, acts through the normal antithrombin mechanism.[15] Figure 4 indicates the important points of this action—namely, that heparin speeds up the rate of normal antithrombin action by a factor of over 50 and at the same time decreases the efficiency of inactivation so that the antithrombin capacity drops by a factor of 1.5 to 2.0. Heparin, a product of the liver, is present in most tissues of the body and can be liberated and appear in the blood stream during anaphylactic shock. The normal concentration of free heparin in the blood, if it is present at all, is so low as to be difficult to measure. From the standpoint of hemostatic feedback mechanisms, the possibility that traces of heparin determine the rate at which normal antithrombin acts is most interesting. Heparin has been mentioned earlier as an important factor in the destruction of plasma thromboplastin.

It is now clear that the shut-off mechanisms attack any active components which leave the clot area. Under more favorable circumstances of time it would be necessary at this juncture to examine some of the concepts concerning the necessity of having threshold levels of procoagulants before the activating mechanisms are effective. There is evidence that such is the case and, if so, threshold levels must be intimately associated with the shut-off mechanism. That is, they would prevent an extensive spread of the explosive development of thrombin under conditions where activators would escape and be carried into the blood stream.

Consider now the clot structure itself. The developing clot has platelets adhering to its fibers, particularly at the fiber junctions. These platelets may be the loci at which such junctions were formed in the first place, it being understood that purified fibrinogens will form firm clots without platelets being present, thus that platelets are not necessary for fibrin development. In any event, the presence of the platelets gives rise in the normal course of events to an expulsion of the fluid portions of the clot, thus to clot retraction. As pointed out at the start, retraction gives rise to a relatively fine-meshed protective wound cover. It is now quite certain that platelets or platelet derivatives (e.g., serotonin) are necessary in the retraction process. The opinion has been expressed that the fibers of the clot are "contracting" in this process of clot retraction. Another alternative, which is the more likely, is that a process of zippering together is initiated at the fiber junctions and that this zippering process extends gradually throughout the clot.

Another important action takes place during clotting. This is the adsorption of an inactive enzyme called profibrinolysin[12] on to the strands of the clot. On activation, profibrinolysin gives rise to the enzyme fibrinolysin whose specificity is directed against fibrin, and to an extent fibrinogen. Fibrin is attacked via the splitting of peptide bonds, the fragments produced being unable to reassociate to form fibrin or a clot structure under the conditions existing in plasma. Not much is known about the normal mechanism for the activation of profibrinolysin, current methods for activation involving such techniques as the shaking of plasma with chloroform or the addition of chemical compounds, for example, streptokinase which is obtained from the corresponding cocci. The focus of attention in this discussion is the fact that fibrinolysin activity is known to be generated within the liquor sanguinis itself: after accidental death, by fear, stress, and severe exercise, after epinephrine injections and at times in "normal" individuals. With an appreciation of the importance of the coagulation mechanisms in everyday physiology, it is not surprising to find an interest in the possibility that a chemical or physical treatment can be found which activates only the fibrinolysin which is adsorbed to the fibrin clot. Another statement, which should come as no surprise, is that there occurs in the blood a potent antifibrinolysin activity which so far has prevented the therapeutic use of fibrinolysin preparations.

At the onset, reference was made to the possibility that a slow, continuous coagulation was necessary in the maintenance of capillary structure. Immediately thereafter, the factors involved in fast clotting and the control of the fast-clotting process were outlined. Return now to the problem of slow clotting, realizing that many of the complications of clotting may occur as a necessity for maintaining and controlling a system acting slowly and continuously, but poised for rapid action.

No absolute proof can be offered in which coagulation factors or the products of coagulation, for example monomer fibrin, can be directly implicated in capillary

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

TABLE II. Hemorrhage owing to clotting defects.

I. Platelet deficiency: from unknown causes or symptomatic	
II. Defects in activating components	
(a) Antihemophilic globulin	(AHG)
(b) Plasma-thromboplastin component	(PTC)
(c) Plasma-thromboplastin antecedent	(PTA)
(d) Accelerator globulin	(AOG)
III. Prothrombin deficiency: lack of vitamin K, caused by drugs, by disease or is congenital	
IV. Fibrinogen deficiency: congenital or through liver damage	
V. Anticoagulants: heparin, enzymes (shock)	

properties such as fragility and permeability. The evidence is indirect and can be summarized briefly by pointing out that interferences in the normal clotting mechanisms often lead to internal hemorrhage, a situation in which blood cells and plasma leak out through the capillaries into the tissue spaces. A few of the coagulation defects which lead to hemorrhage are shown in Table II (see also Wintrobe,[1] p. 285).

First, it is seen that a deficiency in platelets may lead to hemorrhage. This deficiency can arise either through an inherent platelet deficiency or through an external agent, disease for example, which reduces the numbers of platelets.

Second, deficiencies in procoagulation components have the same effect. Here one finds included antihemophilic globulin, plasma-thromboplastin component (mentioned earlier), possibly plasma-thromboplastin antecedent, and accelerator globulin.

Third, a deficiency of prothrombin, whether it be induced by vitamin-K deficiency, by drugs, by disease, or is congenital.

Fourth, a deficiency in fibrinogen: congenital or as a result of liver damage.

Fifth, drugs which increase antithrombin action such as heparin and circulating anticoagulant enzymes which may appear in shock.

The common denominator for this impressive list is active coagulation. In considering the physiological necessity for slow coagulation, the author has been impressed by the difficulty with which the normal anticoagulation mechanisms remove or inhibit the last traces of thrombin. This is a situation so frequently encountered that one is tempted to think in terms of traces of highly resistant thrombin. Such a conclusion would be premature since the alternative that the coagulation system also has a mechanism which preserves slight traces of thrombin is attractive: such a system would not only provide for slow continuous clotting but would supply the traces of thrombin which might spark the initial stage of autocatalytic activation. I feel that the slow-clotting mechanism will emerge as a mechanism warranting close attention in the future.

BIBLIOGRAPHY

[1] M. M. Wintrobe, *Clinical Hematology* (Lea and Febiger, Philadelphia, 1956), fourth edition.

[2] R. Biggs and R. G. MacFarlane, *Human Blood Coagulation and its Disorders* (Charles C. Thomas, Springfield, Illinois, 1957), second edition.

[3] K. M. Brinkhous, R. D. Langdell, and R. H. Wagner, Ann. Rev. Med. **9**, 159 (1958).

[4] W. H. Seegers, Advances in Enzymol. 16, 23 (1955).

[5] K. M. Brinkhous, editor, *International Symposium on Hemophilia and Hemophiloid Diseases* (University of North Carolina Press, Chapel Hill, North Carolina, 1957).

[6] W. H. Seegers, R. I. McClaughry, and J. L. Fahey, Blood **5**, 421 (1950).

[7] F. Lamy and D. F. Waugh, Physiol. Rev. **34**, 722 (1954).

[8] S. J. Shulman, J. Am. Chem. Soc. **74**, 5706 (1952).

[9] J. E. Fitzgerald, N. S. Schneider, and D. F. Waugh, J. Am. Chem. Soc. **79**, 601 (1957).

[10] L. Lorand and A. Jacobsen, J. Biol. Chem. **230**, 421 (1958).

[11] W. A. Scheraga and M. Laskowski, Jr., Advances in Protein Chem. **12**, 1 (1958).

[12] S. Shulman, N. Alkjaersig, and S. Sherry, J. Biol. Chem. **233**, 91 (1958).

[13] W. H. Seegers, B. H. Landaburu, R. R. Holburn, and L. M. Tocantins, Proc. Soc. Exptl. Biol. Med. **95**, 583 (1957).

[14] D. F. Waugh and B. J. Livingstone, J. Phys. & Colloid Chem. **55**, 1206 (1951).

[15] D. F. Waugh and M. A. Fitzgerald, Am. J. Physiol. **14**, 627 (1956).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

61
Hormone Regulation

DEWITT STETTEN, JR.

National Institute of Arthritis and Metabolic Diseases, National Institutes of Health, Public Health Service, U. S. Department of Health, Education, and Welfare, Bethesda 14, Maryland

IN higher organisms, whether animal or vegetable, there exist channels through which fluids are transmitted from one tissue to another. Into these fluids or humors are delivered many compounds elaborated in one or another tissue. It is to be expected that, in many cases, these compounds will exert recognizable physiological effects in other tissues to which they may be hydraulically transmitted. Among the host of compounds which meet these criteria, a limited number have been set apart under the special category of hormone or "chemical messenger." It is the purpose here to review briefly what these hormones are, whence they arise, whither they go, and how they accomplish their missions. Although hormonal mechanisms are well described in invertebrates and indeed in plants, these remarks are restricted to mammalian endocrinology.

To speak of the mission of a compound is to lapse into teleological jargon. Teleology, however, has been defined as a woman with whom no scientist cares to be seen in public, but with whom all scientists consort in private. It will simplify my task if you will forgive my teleological indiscretions in this paper which permits me to dodge the otherwise necessary cumbersome circumlocutions.

What then is a hormone? It is a compound devoid of intelligence but containing nonetheless in its structure certain information. While there are no Koch's postulates by which one may judge whether a compound is or is not a hormone, there are certain accepted standards to which most respectable hormones conform.

1. In many but not all cases, hormones arise in tissues which appear to be specialized for their production. In the mammal, these organs include the neurohypophysis, the adenohypophysis, the thyroid, the parathyroids, the islets of Langerhans of the pancreas, the adrenal medullae, the adrenal cortices, the testes, the ovarian follicles, and the ovarian corpora lutea. Other organs, such as the pineal body, the hypothalamus, the thymus, the carotid bodies, etc., are not included in the present list of so-called endocrine organs in view of their relatively uncertain status and lack of specific and convincing information.

It should not be supposed, however, that all known hormones arise in tissues anatomically specialized for their production. The word "hormone" was coined actually to describe an as yet unidentified principle termed secretin, discovered by Bayliss and Starling, which appears to arise in the intestinal mucosa, to enter the blood stream, and to stimulate, upon its arrival there, the secretion of juices by the pancreas. A second but distinct material, arising from the same source, cholecystokinin, stimulates the contracture of the gall bladder, and the injection of bile into the duodenal lumen. Similarly, there is partial evidence for the existence of renal hormones affecting vascular tone, hepatic hormones affecting uptake of blood glucose by muscle, and indeed a hormone liberated by working skeletal muscle which affects the metabolism of resting muscle in the same animal.

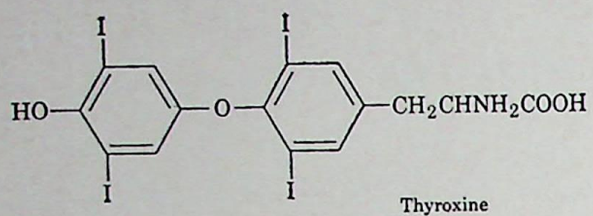
Whereas a number of the hormones which arise in the specialized endocrine glands, e.g., the pituitary, the adrenal, etc., have been purified and characterized structurally, this has not been accomplished for hormones arising from such tissues as liver or intestinal mucosa. Consequently, the structural information presented is confined to products of the so-called glands of internal secretion.

2. A second characteristic of hormones is that they normally circulate and exert their regulatory roles at low molar concentrations. On the basis of published estimates of the normal concentration of insulin in blood, it would appear that this agent is effective at concentrations of about 10^{-11} to 10^{-12} molar, and epinephrine, glucagon, and thyroxin all appear to exert recognizable effects at similar levels of concentration.

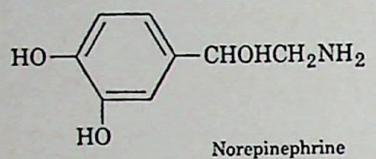
3. Some, but not all, hormones exert their effects upon highly specific target organs. Thus, adrenocorticotropic hormone of the pituitary operates essentially uniquely upon the adrenal cortex, while the activity of thyrotropic hormone is restricted largely to the thyroid. The sole convincingly demonstrated action of glucagon is upon the liver. On the other hand, epinephrine exerts actions upon liver and muscle, insulin has been shown to operate in many tissues including skeletal muscle, adipose tissue, and mammary gland, whereas the effects of adrenocortical steroids and thyroxin may be demonstrated in a wide variety of tissues.

4. If a hormone is to exert a regulatory effect, clearly the rate of its release into the blood stream must be variable and should be determined by some function of the mechanism which it regulates. Such variation in rate of discharge has been demonstrated for certain but not for all hormones. For example, one may cite insulin, which regulates the level of glucose in the blood and the release of which by the islet cells of the pancreas is determined directly by the concentration of glucose in the pancreatic blood supply. Analogously, ACTH, released by the pituitary, specifically stimulates the adrenal cortex to discharge steroid hormones, which, if injected, cause a prompt suppression of ACTH produc-

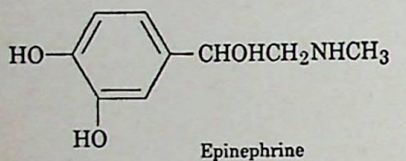
CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA



Thyroxine



Norepinephrine



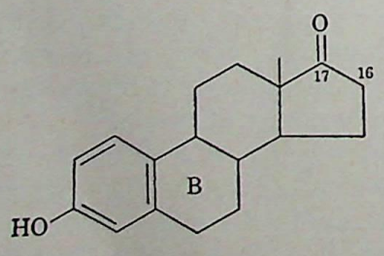
Epinephrine

FIG. 1. Structures of certain phenolic hormones.

tion. Many other such examples of negative feedback, more or less well-documented, might be cited.

It may be profitable to consider briefly some of the structural types which are represented among the hormones. These include a variety of substances, among the simplest of which are certain phenols (Fig. 1). Under the regulation of thyrotropic hormone derived from the anterior pituitary gland, a protein, thyroglobulin, is elaborated in the thyroid gland, which contains the unique amino acid thyroxin, 3,5,3′,5′-tetraiodothyronine. This amino acid, accompanied by triiodothyronine, is discharged into the bloodstream where it circulates, largely protein-bound, in the plasma. Epinephrine and norepinephrine derived from adrenal medulla are related closely both metabolically and structurally. There is a considerable amount of information at hand defining the biosynthetic origins of these compounds. Tyrosine contributes the phenolic portions, in each case. With these, as well as with certain other hormones, secondarysites of origin exist. The synthesis of thyroxin-like materials in the totally thyroidectomized animal[1] is perhaps not surprising in view of the fact that tyrosine-containing proteins, after incubation with iodine in alkali, in the absence of enzymes, exhibit thyroxin activity.

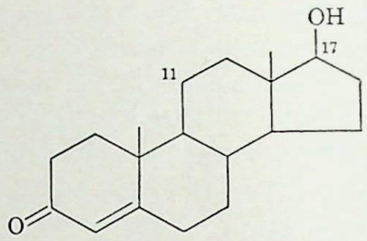
Steroidal hormones comprise a large and much-studied group. They often are classified structurally according to the number of carbon atoms contained. Thus, the C_{18} steroids include the estrogens (Fig. 2), estrone, α- and β-estradiol, estriol, as well as the less familiar equilin and equilenin. All of these compounds have phenolic—i.e., acidic—properties. The methyl group born on carbon atom 10 of most other steroids is lacking in these compounds, as a consequence of the unsaturation. Arising predominantly in the Graafian follicle of the ovary, under regulation by pituitary gonadotropin, estrogens are generated also by testis and possibly also by adrenal cortex. They are detoxified normally in the liver by conjugation with glucuronic or sulfuric acids, and the loss of this function in hepatic failure results in feminizing symptoms and signs.



Estrone

Substituent at C Numbers		
Compound	17	16
α or β Estradiol	—OH	—H
Estriol	—OH	—OH

FIG. 2. Structures of certain C_{18} steroid hormones (N. B. Equilin and equilenin represent successive states of oxidation of B ring of estrone).



Testosterone

Substituent at C Numbers			
Compound	3	11	17
Adrenosterone	=O	=O	=O
Androstenedione	=O	—H	=O
Androstanediol*	—OH	—OH	=O

* Double bond at Δ^4 is reduced

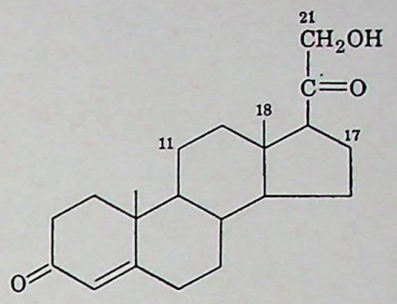
FIG. 3. Structures of certain C_{19} steroid hormones.

The C_{19} steroids include (Fig. 3) the androgens, testosterone and its congeners. There are a number of steroids with androgenic activity and these may arise in the adrenal cortex as well as in the testis. The C_{21} steroids (Fig. 4) include not only the several adrenocortical steroids but also progesterone, the hormone of the ovarian corpus luteum. The generation of these hormones is under direct control of specific anterior pituitary hormones. Testicular production of androgens is regulated by pituitary gonadotropin, adrenocortical activity is determined by pituitary adrenocorticotropin, and progesterone production depends upon a luteinizing hormone of the pituitary gland.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

The structural similarities of the several steroid hormones are explained by the fact that they are all biogenetically closely related. Cholesterol appears to be a parent compound to all of them, and many of the individual enzyme-catalyzed steps whereby they arise have been elucidated. It is, of course, of great interest that specific and important biological activities of a wide variety are associated with steroids. In addition to the steroid hormones, other biologically important steroids are vitamin D, the cardiac aglycones of the digitalis group of drugs, the toad poisons, and certain carcinogens. What it is in the steroid structure which imparts to the members of this class these various activities is not known.

The remaining groups of hormones about which there is structural information are polypeptides. Of these, insulin (Fig. 5) is probably the best-studied. Beef insulin, whose chemical structure was given by Sanger,[2] is made up of two polypeptides, an *A*- or acidic chain comprising 21 amino acids and a *B*- or basic chain containing 29 amino acids. The intramolecular disulfide bridges have been described previously by Oncley (p. 30), and their integrity is essential to the physiological activity. The C-terminal alanine in the *B*-chain is apparently not essential, however, and indeed the eight C-terminal amino acids of this chain may be eliminated without complete inactivation. The question of what constitutes the essential center of physiological activity of insulin, if indeed there is such, remains unanswered. When this structure first was published, attention was attracted to the pentapeptide closed ring in the *A*-chain, particularly since rings of similar size were found by du Vigneaud[3,4] to be present in both of the hormones of the posterior pituitary gland, vasopressin and oxytocin. The curious run of three adjacent aromatic amino acids in the *B*-chain also was pointed out.



Deoxycorticosterone

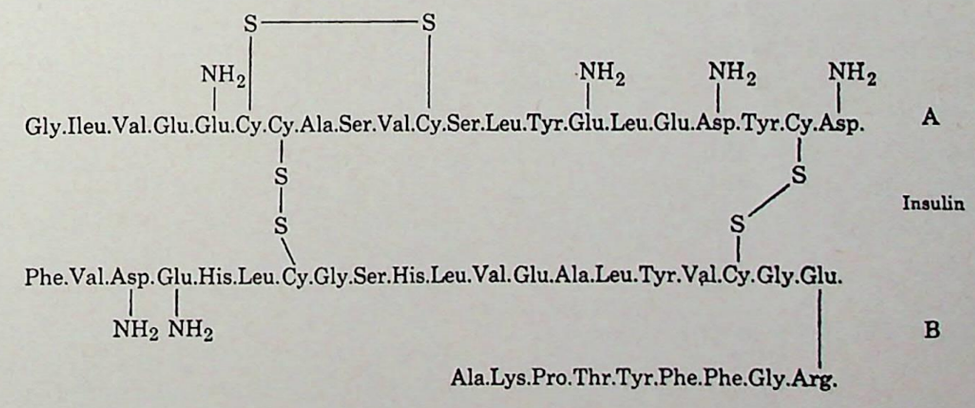
Compound	Substituent at C Numbers 11	17	18
Corticosterone	—OH	—H	—CH_3
11-Dehydrocorticosterone	=O	—H	—CH_3
17-Hydroxycorticosterone	—OH	—OH	—CH_3
Cortisone	=O	—OH	—CH_3
Aldosterone	—OH	—H	—CHO

FIG. 4. Structures of certain C_{21} steroid hormones (N. B. Progesterone differs from deoxycorticosterone only in that $-CH_2OH(21)$ is reduced to $-CH_3$.

No evidence has been presented implicating either of these sites specifically in insulin function.

The posterior pituitary hormones present striking similarities and, as might be expected, some degree of cross-reactivity physiologically. There is some reason to believe that these compounds may be parts of a larger polypeptide arising in the hypothalamus and transmitted down the pituitary stalk, to be stored in and released from the posterior lobe of the pituitary gland.

Attention is directed now to amino acids 8, 9, and 10 in the *A*-chain of insulin. Insulins of several species have been prepared; all are very similar in physical properties and are essentially identical in physiological assay, about 27 units per mg. Chemical differences do exist,



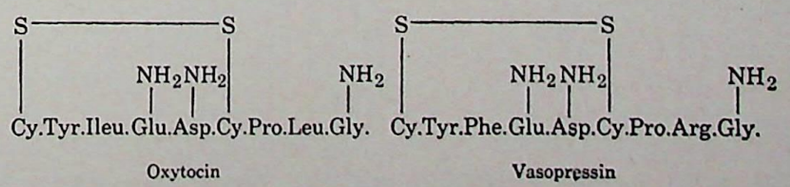


FIG. 5. Structures of bovine insulin and the hormones of the posterior pituitary.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

Insulin–A–Chain*

1 — — 7 8 9 10 — — — — 20

...Cys.Ala.Ser.Val...	Beef
...Cys.Thr.Ser.Ileu..	Pig †
...Cys.Ala.Gly.Val...	Sheep
...Cys.Thr.Gly.Ileu..	Horse
...Cys.Thr.Ser.Ileu..	Whale †

*Remainder of A chain and entire B chain (1 through 29) are invariant

FIG. 6. Species variations in insulin structure.

however, but thus far the sole differences that have been detected reside within these 3 amino acids (Fig. 6). Although the sample studied is confessedly too small to permit firm conclusions to be drawn, a few suggestive findings have turned up. In each position, 8, 9, or 10, only one of 2 amino acids has been found. Thus, in position 8, there is either alanine or threonine; in position 9, serine or glycine; in position 10, valine or isoleucine. If this binary replacement proves general, and no other alterations are found, only 8 insulins are possible. There may be, however, an additional restriction. On comparison of positions 8 and 10, it appears that, whenever the former is alanine, the latter is valine; whenever the former is threonine, the latter is isoleucine. Should this represent a firm link in the coding to these 2 positions, only 4 insulins would occur, and indeed all types now are known. It is not surprising, then, that pig and whale insulins prove to be identical (Fig. 6†).

Insulin is generated in the β cells of the islets of Langerhans. Its function, as is discussed later, is to favor the utilization of blood glucose by muscle, to restrict glucose formation by liver. It is noteworthy, therefore, that the release of insulin by β cells is favored by a high concentration of blood glucose.

Intimately associated anatomically with the β cells are the α cells of the islets which are the source of glucagon. This polypeptide (Fig. 7) was for many years an unsuspected contaminant of the best available insulin from American commercial sources although, because of differences in manufacturing procedure, it was lacking in Danish insulin. Chemically and physiologically it is

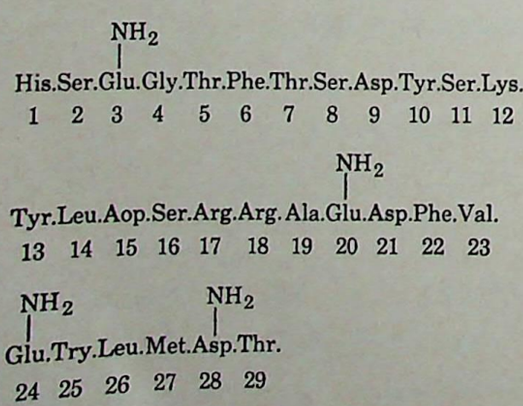


FIG. 7. Structure of glucagon.

unrelated to insulin.[5] The sole sequential similarity between insulin and glucagon that has been pointed out is the series 22–24 in glucagon (Phe. Val. Glu NH_2) which resembles Phe. Val. Asp NH_2 (1–3 chain B) of insulin. Whereas both glucagon and insulin affect the level of blood glucose, they do so by influencing quite different processes, and appear to be antagonists only on very superficial inspection.

The last group of hormones about which there is useful structural information are the polypeptides of the anterior pituitary gland. There are at least six of these, thyrotropin, adrenocorticotropin, somatotropin (growth hormone), follicle stimulating, luteinizing, and melanophore stimulating hormones. Other activities have been suspected in the past and may be revealed in the future. Complete structural information is available, for certain species, in regard to corticotropin and melanotropins. The formulas for certain of these have been given by Oncley (p. 30). Here, attention is directed only to certain salient points of similarity and dissimilarity (Fig. 8). Amino acids 4 through 10 of all corticotropins studied are identical with amino acids 7 through 13 of all known melanotropins. This identity of a series of seven amino acids is sufficiently startling

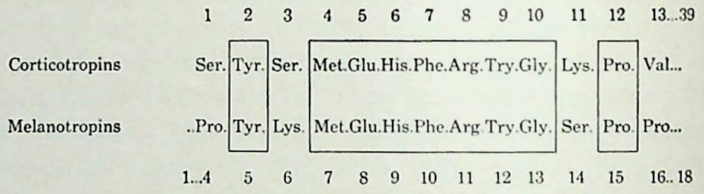


FIG. 8. Structural relationships of ACTH and melanophore stimulating hormone.

to attract attention. Li[6] has pointed out that the identical sequences would have been eleven rather than seven, but for a curious confusion about positions 3 and 11 of corticotropins, or 6 and 14 in melanotropins. An interchange between the serine and lysine in the appropriate positions of either of these peptides would prolong the identical sequences from seven to eleven.

This observation, as well as the previously mentioned species variations in insulin, suggests the occurrence of certain blunders in the transmission of the coded information pertaining to amino-acid sequences. Such blunders, if indeed so they be, may themselves be informative. Thus, the present figure suggests that the code is such as to permit interchange of information about 2 amino acids 8 positions remote in a polypeptide.

To date, accumulating knowledge of hormone structure has, for the most part, revealed little about how these compounds function. Only within the past few years have reasonable hormone effects been observed in biological preparations disorganized below the cellular level. Several of these now may be mentioned.

Thyroxin has been suspected for some time of exerting an effect similar to that of 2,5-dinitrophenol, a suspicion dating back to the early clinical observations of the

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

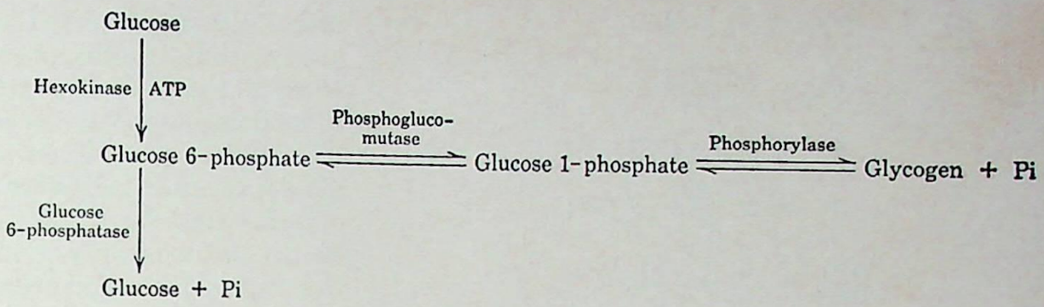


FIG. 9. Certain phases of the metabolism of glucose.

toxicology of the latter drug during World War I. When it was ascertained that dinitrophenol effected an uncoupling of oxidative phosphorylation, permitting respiration to proceed without obligatory and coincident phosphorylation, a similar effect was attributed to, and soon found with, thyroxin. This effect has been demonstrated with isolated mitochondria by Lehninger.[7] Such uncoupling, and consequent escape of glycolysis from the restraining influence of the Pasteur effect, would explain many of the clinical manifestations of thyrotoxicosis.

Estrogens have been shown by Villee[8] to activate the isocitric-acid dehydrogenase of placenta and this effect has been demonstrated in homogenized preparations (cell-free). More recently, Talalay[9] has indicated that the enzyme of placenta directly affected is a transhydrogenase which regenerates the TPNH required by the dehydrogenase in question. It is possible that, in this transhydrogenase reaction, the estrogen serves as a self-regenerating co-substrate.

Two hormones about which there is partial mechanistic information are epinephrine and glucagon. It is perhaps remarkable that these two substances, wholly unrelated structurally and biogenetically, should both exert, at equimolar concentrations, virtually identical effects. These effects, elucidated by Sutherland,[10] relate to the activation of hepatic phosphorylase. This enzyme, required for glycogen breakdown and probably also for its synthesis, is itself enzymatically inactivated with the loss of inorganic phosphate. Its reactivation also may be accomplished enzymatically, but this requires a novel cofactor, an anhydride of AMP. The synthesis of this cofactor, by yet another enzyme, proceeds only in the presence of epinephrine and/or glucagon. The two hormones, though acting identically in liver, are quite different in their effects upon muscle glycogen. Here, whereas epinephrine is highly active, glucagon is without effect.

Insulin, which has received the attention of many excellent investigators since its isolation, is in a more confused state than some of the other hormones which have been considered. The dramatic capacity to lower blood sugar apparently is the result of effects upon two processes. To evaluate these two effects, one should consider briefly some of the sources and fates of blood glucose (Fig. 9). Here are represented in abbreviated form some of the reactions of concern. In the absence of insulin, as in the severely diabetic human or as in the pancreatectomized animal, the conversion of extracellular glucose into intracellular glucose 6-phosphate, an essential first step in glucose utilization, is impaired. This impairment is especially demonstrable in muscle tissue and characteristically is seen in the isolated diaphragm. The basis for the impairment is still somewhat obscure but the bulk of current evidence favors the view, first propounded by Levine,[11] that insulin favors the entry of glucose into the intracellular compartment of muscle. Insulin demonstrably performs this function in relation to a number of other sugars. Possibly significant is the fact that the sugars most responsive to insulin in this regard are configurationally identical with glucose about carbon atoms 1, 2, and 3.

A second pertinent insulin effect is defined best by the fact that the activity of glucose 6-phosphatase is greatly enhanced in the liver of the diabetic animal, an effect reversible by administration of insulin. The importance of this enzyme is revealed by the fact that it catalyzes the reaction responsible for the majority of glucose contributed to the blood by animal tissues.

Consider the dilemma of the diabetic animal in the light of these two defects. With the conversion of glucose to glucose 6-phosphate restricted, and with the hydrolysis of this ester exaggerated, there results a curtailment of glucose phosphate needed for a variety of important processes, including glycolysis, oxidation, generation of reduced coenzymes, preparation of essential intermediates. Many, but certainly not all, of the late consequences of untreated diabetes are referable directly to this combination of biochemical defects. The question of which of the two defects is most significant is in a state of flux at the present time. Data from Hastings[12] and others suggest that the increase in utilization of glucose by muscle occurs much more promptly in response to insulin than does a decrease in glucose 6-phosphatase activity. On the other hand, Weinhouse and his collaborators[13] have data indicating that the reverse is true.

In relation to this figure, one can consider also the purported "antagonism" between insulin and glucagon. Glucagon, as mentioned previously, effects favorably the activity of hepatic phosphorylase, promotes breakdown of liver glycogen, and produces a resulting rise in blood glucose. Insulin, by its actions, produces a fall in blood glucose. If one's vision is restricted to the glucose of the blood, these two hormones act antagonistically. The blood glucose, however, is really relatively un-

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

$$HT \rightleftharpoons H + T$$

$$K = \frac{[H]\,[T]}{[HT]}$$

$$Q = [T] + [HT]$$

$$Q = K\frac{[HT]}{[H]} + [HT]$$

$$\frac{[H]}{[HT]} = \frac{K}{Q} + \frac{[H]}{Q}$$

or

$$\frac{1}{[HT]} = \frac{1}{[H]} \cdot \frac{K}{Q} + \frac{1}{Q}$$

FIG. 10. Expected arithmetical relationship between concentrations of free hormone, H, of bound hormone, HT, and of acceptor sites on target organ, Q.

important, except insofar as it nourishes the several tissues of the body, and the major consuming tissue is skeletal muscle. With respect to the nutrition of skeletal muscle, glucagon and insulin, far from being antagonists, are synergists. Glucagon supplies to the blood the glucose which muscle requires, and insulin facilitates the entry of this glucose into the cells so that it may be utilized.

It was mentioned at the outset that many endocrines, insulin among these, are physiologically active at remarkably low concentrations. The finding of Stadie[14] that several tissues responding to insulin *in vitro* are endowed with the capacity to bind insulin to their surfaces, and that the response to insulin is, under appropriate circumstances, proportional to the amount of insulin bound, is perhaps relevant. If the target organ has the capacity to bind the hormone to itself, local concentrations of hormone may be built up far in excess of the concentrations occurring in the blood. The dose-response relationship for several hormones is compatible with this idea in that, whereas the response is approximately proportional to dosage at low dosage levels, for many hormones a maximal response is attained asymptotically at high dosage levels. This suggests that something, possibly the acceptor site on the target organ, is being saturated with hormone. This situation may be treated arithmetically[15] by assuming a reversible association between hormone (H) and target-organ acceptor site (T). If K is the dissociation constant and Q is the total concentration of acceptor sites (Fig. 10), the problem may be treated as Lineweaver and Burk treated the Michaelis-Menten problem, to yield the final expression on the figure. If now it be assumed tentatively that the concentration of hormone H is proportional to dosage and that the response is proportional to the concentration of bound hormone HT, the reciprocals of these two quantities should be related linearly. In a limited number of cases where this relationship has been tested, the expected linear relationship indeed has been found. Values for K and Q can be estimated from the slopes and intercepts of such plots. The usefulness of the relationship diminishes insofar as some quantity other than the concentration of bound hormone limits the observed response.

It is perhaps unfortunate to terminate this distinguished series upon such an unsatisfactory note. In earlier papers on macromolecular structure and interaction, biological fine structure, information storage and retrieval, etc., one was left with the idea that there is a narrowing gap between the biological description and the physical model, or even explanation. In the matters presently under consideration, the gap is so wide that I, for one, am unable to see across it. Of such a gap, one may hope, but one can not know, that it is being narrowed. Certainly many experiments will have to be performed before the hormones, their origins, their modes of action, and their interrelationships can be defined in physical terms.

BIBLIOGRAPHY

[1] M. E. Morton, I. L. Chaikoff, W. O. Reinhardt, and E. Anderson, J. Biol. Chem. **147**, 757 (1943).

[2] F. Sanger, L. F. Smith, and R. Kitai, Biochem. J. **58**, vi (1954).

[3] V. du Vigneaud, C. Ressler, and S. Trippet, J. Biol. Chem. **205**, 949 (1953).

[4] V. du Vigneaud, H. C. Lawler, and E. A. Popenoe, J. Am. Chem. Soc. **75**, 4880 (1953).

[5] W. W. Brommer, L. G. Sinn, A. Staub, and O. K. Behrens, J. Am. Chem. Soc. **78**, 3858 (1956).

[6] C. H. Li, Advances in Protein Chem. **12**, 269 (1957).

[7] A. L. Lehninger in *Enzymes: Units of Biological Structure and Function*, O. H. Gaebler, editor (Academic Press, Inc., New York, 1956), p. 217.

[8] C. A. Villee and D. D. Hagerman, Endocrinology **60**, 552 (1957).

[9] P. Talalay and H. G. Williams-Ashman, Proc. Natl. Acad. Sci. U.S. **44**, 15 (1958).

[10] T. W. Rall, E. W. Sutherland, and J. Berthet, J. Biol. Chem. **224**, 463 (1957).

[11] R. Levine, M. S. Goldstein, B. Huddlestun, and S. F. Klein, Am. J. Physiol. **163**, 70 (1950).

[12] A. E. Renold, A. B. Hastings, F. B. Nesbett, and J. Ashmore, J. Biol. Chem. **213**, 135 (1955).

[13] D. F. Dunn, B. Friedmann, A. R. Maass, G. A. Reichard, and S. Weinhouse, J. Biol. Chem. **225**, 225 (1957).

[14] W. C. Stadie, N. Haugaard, and M. Vaughan, J. Biol. Chem. **200**, 745 (1953).

[15] DeW. Stetten, Jr., Science **124**, 365 (1956).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

REVIEWS OF
MODERN PHYSICS

VOLUME 31, NUMBER 3 JULY, 1959

Structure of Spectral Lines from Plasmas*

HENRY MARGENAU AND MARVIN LEWIS
Yale University, New Haven, Connecticut

CONTENTS

I. Introduction 569
A. Statistical Theory 570
B. Impact Theory 571
C. Relation between Statistical and Impact Theories 572
II. Classical-Path Hypothesis and Criteria for Adiabaticity 575
A. Classical Path 575
B. Criteria for Adiabaticity 578
C. Construction of Wave Packets 581
D. Summary and Appraisal 581
III. Holtsmark Theory 582
A. Review of Conventional Treatment . . . 582
B. Effect of Perturber Interactions 584
C. Treatment for Radiating Ions 585
IV. Refined Impact Theory. Fusion of Impact and Statistical Theories 588
A. Impact Theory 588
B. Core and Wing Theorems for Individual Collisions 591
C. Some Numerical Estimates, Mainly Regarding the Balmer Lines 593
V. Disappearance of Lines with High Quantum Numbers 594
VI. Broadening of Degenerate Levels 596
A. Mathematical Foundations 597
B. Adiabatic Approximation 598
C. Diabatic Effects in Degenerate Systems 599
D. Electron Broadening in the Adiabatic Approximation 600
E. Broadening of the Lyman-α Line by Electrons 601
F. Simultaneous Broadening by Ions and Electrons, in the Adiabatic Approximation 601
VII. Experimental Results and Classical Interpretations 603
A. Experimental Determination of Plasma Properties 603
B. Some Recent Experimental Results 605
C. Classical Interpretations 606
VIII. Quantum Treatment of the Lyman-α Line 608
IX. Quantum Theory of the Balmer Lines 611
X. Shifts Produced by Electron Collisions 613
A. Quantum Theory 614
B. Application to Helium 615
C. Comparison with Lindholm's Theory . . 615

I. INTRODUCTION

RESEARCH on line broadening, though often regarded as pedestrian and unlikely to lead to new fundamental insights, is nonetheless inspired by a vision. One dreams that in a distant star or in some other inaccessible region filled with matter a few atoms, perhaps hydrogen atoms, emit lines whose structure can be analyzed in terrestrial laboratories. These lines carry in many cases information we have already learned to understand: red shifts revealing masses, Doppler shifts revealing motions, and sometimes Doppler widths revealing temperatures. It is clear that in principle all physical properties of the medium containing the radiating atoms are somehow reflected in the line structure, since they affect the forces between the radiating atom and its neighbors, the distances over which these forces are exerted, and the times during which they act. One hopes, therefore, that when the language of the spectral lines has been fully learned, a radiating atom in a distant material environment can serve as a noninterfering probe conveying significant data regarding pressure, temperature, or, more generally, the distribution of molecular speeds, and states of ionization in the surrounding medium. Such hope is now far from fulfillment, and this article attempts but a modest contribution to its realization.

* Work supported by the Air Force Office of Scientific Research.

Copyright © 1959 by the American Physical Society
CC-0. Gurukul Kangri University Haridwar Collection. Digitized by eGangotri Foundation USA

This survey is limited to a small part of the line-width problem, a problem whose scope and complexity is not widely appreciated and which has suffered perhaps from an excessive optimism of authors who, on proposing a partially successful simple model yielding to mathematical treatment, have served enthusiastic notices that the whole problem is solved. When, as often happens, experimental observations fit the special model, the matter is easily regarded as closed despite logical difficulties, and it sometimes takes severely critical appraisals to set the problem on the path of progress again.

The part treated here is plasma broadening, a group of effects peculiar to lines emitted in a medium that is strongly ionized but has no net charge. Hence we do not consider the results of impacts between neutral molecules, nor any of the theories suitable principally for their description. Other review[1] articles fill this need. In a plasma, the heavy positive charges and the light electrons often require different treatment, as the following discussion shows. Attention is restricted to emission because in ionized media absorption is not often directly observable and the problems connected with phenomena of reabsorption necessitate considerations unrelated to the broadening agencies here reviewed. Nor are microwaves specifically included in our work; optical lines are our chief concern.

Applications of the theory are found in four widely separated fields of investigation: in astrophysics, gas discharges, strong shocks, and flames or explosions. Some practical urgency attaches to line-width studies because of the importance which all of these researches currently enjoy.

This introductory section surveys two very general methods that have found application to plasma line-broadening problems. One is called statistical theory; the other is variously called impact theory, velocity broadening or phase-shift broadening theory. They refer to different models that cannot easily be compared. Because of the intuitive difference of the models the effects they generate are sometimes regarded as separate or even additive, and oversimplifications have appeared as a result of this seductive fallacy. Statistical and impact theories are certain limiting instances of a more general theory, each having its own range of application, and each losing its validity in a large domain of physical conditions of interest to the experimenter. There are also situations in which both theories apply and give identical answers. But it happens that in plasma broadening the integrity of the two methods is rather better preserved than under almost any other circumstances, the reason being that the heavy ions exert long-range and slowly varying forces which satisfy the statistical theory, whereas the light and swift electrons are very often tractable by impact methods.

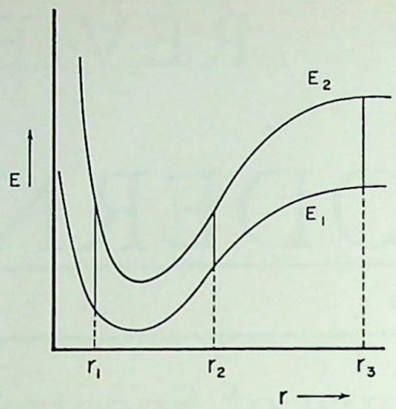


FIG. 1. Dependence of energies on interparticle distance.

A. Statistical Theory[2–4]

A spectral line arising from a transition between an upper level of energy E_2 and a lower of energy E_1, has its normal (radian) frequency $(E_2-E_1)/\hbar$. But E_2 and E_1 are not constants; they vary as a result of perturbations caused by other molecules, the "perturbers." In the presence of only one perturber the dependence of E_2 and E_1 on the distance, r, between the centers of the radiating and perturbing particle is given by the well-known potential energy curves, of which a typical set is drawn in Fig. 1. A transition occurring at a large distance r_3 has a normal ω, at r_2, ω is smaller and at r_1 larger than normal. Thus to every r there corresponds an $\omega(r)\equiv(E_2(r)-E_1(r))/\hbar$ and some ω's are more likely than others. In a statistical sense, the probability $P(\omega)$ equals the fraction of configuration space in which $E_2-E_1=\hbar\omega$ or, more precisely, that fraction times the Boltzmann factor, $\exp\{-[E_2(r)-E_2(\infty)]/kT\}$. It has been customary to neglect the Boltzmann factor on the supposition that it is practically 1 where $P(\omega)$ is large. This may sometimes lead to serious errors; yet we assume it here.

In the presence of N perturbers, the frequency distribution $I(\omega)$ is still given by $P(\omega)$, but configuration space now has $3N$ dimensions, and

$$I(\omega)d\omega=P(\omega)d(\omega)$$

$$=(4\pi/V)^N\int\cdots\int r_1^2r_2^2\cdots r_N^2dr_1\cdots dr_N. \quad (1.1)$$

The integration extends over the restricted domain in which

$$\omega-d\omega\leqslant\omega(r_1,r_2\cdots r_N)\leqslant\omega+d\omega.$$

[1] (a) V. Weisskopf, Physik. Z. 34, 1 (1933); (b) H. Margenau and W. W. Watson, Revs. Modern Phys. 8, 22 (1936); (c) A. Unsöld, Vierteljahresschr. astron. Ges. 78, 213 (1943); (d) S. Ch'en and M. Takeo, Revs. Modern Phys. 29, 20 (1957); (e) R. G. Breene, Revs. Modern Phys. 29, 94, 1957; (f) I. I. Sobel'man, Progr. Phys. Sci. (U.S.S.R.) 54, 551 (1954). (This paper contains useful comments on the interplay between the Doppler effect and interaction broadening, a problem not considered here.)

[2] (a) H. Margenau, Phys. Rev. 40, 387 (1932); (b) 43, 129 (1933); (c) 44, 931 (1933); (d) 48, 755 (1935); (e) 82, 156 (1951).

[3] M. Kulp, Z. Physik 79, 495 (1932); 87, 245 (1933).

[4] (a) H. Kuhn, Phil. Mag. 18, 987 (1934); (b) H. Kuhn and F. London, Phil. Mag. 18, 983 (1934).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

Usually, in application of statistical theory, additivity is assumed for the contributions to ω:

$$\omega(r_1 \cdots r_N) = \sum_i \omega(r_i).$$

This is often true for large r, but never for small r, where the main contribution to the interaction comes from the exchange forces. In a plasma, $\hbar\omega(r)$ for large r is the difference in atomic energy resulting from the Coulomb forces which are exerted by the passing ions and electrons. Mathematical techniques for evaluating the integrals in (1.1) involve the use of δ functions (Dirichlet's method) now so common as not to require review (see Margenau[2c] and Chandrasekhar[5]).

B. Impact Theory

An infinitely sharp line implies a radiating process of infinite duration. Natural line width arises from the fact that the upper state of the atom has a finite life time, and can be obtained by carrying out a Fourier analysis of the sinusoidal vibration for a finite time. According to the reasoning of Lorentz,[6] impacts of perturbers shorten the lifetime below the natural duration and thereby increase the width of the emitted line. Mathematical representation of this process is very simple: Suppose an atom is allowed to emit radiation of frequency ω' between times 0 and T. Then the distribution of Fourier amplitudes is given by

$$J(\omega,T) \propto \int_0^T e^{i(\omega-\omega')t} dt = \frac{e^{i(\omega-\omega')T}-1}{i(\omega-\omega')}.$$

If this were the only radiative act observed, the frequency distribution would be $|J(\omega,T)|^2$. The observed line, however, is a composite of radiations from many atoms radiating for different lengths of time, but with a mean time equal to τ, the reciprocal of the collision frequency ν_c. The probability that a given atom radiate for T seconds is $\tau^{-1}e^{-T/\tau}$. Hence the intensity at ω is

$$I(\omega) = \tau^{-1}\int_0^\infty |J(\omega,T)|^2 e^{-T/\tau} dT \propto \frac{1}{(\omega-\omega')^2+\nu_c^2}. \quad (1.2)$$

This is the famous Lorentz "dispersion" curve whose full width at half-maximum is $2\nu_c$ on the radian frequency scale. In this article the "half-width" $\omega_{\frac{1}{2}}$ is defined as full width of the line at a height equal to $\frac{1}{2}$ its maximum intensity. Formula (1.2), despite its simplicity, meets with singular success in predicting that a "pressure-broadened" line has a width proportional to the collision frequency, and since

$$\nu_c = nqv \quad (1.3)$$

in terms of the number density of perturbers n, the collision cross section q, and their velocity v, this width turns out to be proportional to the pressure at a constant temperature. Hence the name, pressure broadening.

The meaning of an impact is clear only if the molecules are rigid spheres, an hypothesis of the worst sort for ions. Even for neutral structures, Eq. (1.2) is not wholly satisfactory, for if used in conjunction with (1.3) q must take on values quite different from gas-kinetic cross sections, usually larger values. To hide the difficulty a new phrase was coined; theorists began to speak of *optical* cross sections, to the delight of experimenters who now had a brand new quantity to measure. The unpleasant fact is, however, that Lorentz' theory in its simple form breaks down when the perturbers are surrounded by force fields, i.e., when impacts are "soft".

Qualitatively speaking, collisions affect the radiative process in two ways:

(1) Some actually terminate it, quenching the upper state by transferring the energy of excitation prematurely to another place. Precisely under what conditions an impact will quench is not easy to say because this process requires a delicate balance of energies within perturber and radiator. But it is known that quenching occurs, because it manifests itself in an added line width *and* a reduction in the total line intensity.

(2) The other collision effect is a phase change in the emitted radiation accompanying the detuning of its frequency by a passing particle. If the phase change connected with a collision is large enough the passage will effectively divide the wave train into two incoherent ones, and this is tantamount to termination so far as line width is concerned. In this case no change in intensity accompanies this interruption because the atom goes on radiating a similar line during each successive free period. Lorentz did not distinguish between quenching and phase-altering impacts.

Lenz and Weisskopf[7] incorporated the idea of phase changes in the impact theory and achieved thereby a measure of success in giving meaning to an optical cross section, or an "optical impact radius." In the absence of perturbations, Weisskopf reasoned, ω=constant=ω' and the phase is a linear function of t, $\omega' t$. This corresponds to a sharp line. If an appreciable change $\Delta\varphi$ is added to this linear trend during a perturbation, that perturbation acts like Lorentz's impact. Now

$$\Delta\varphi = \int(\omega-\omega')dt = \hbar^{-1}\int(\Delta E_2 - \Delta E_1)dt, \quad (1.4)$$

where $\Delta E(r)$ is the departure of the E curves in Fig. 1 from their final values. Writing $\epsilon(r)$ for the second difference $\Delta(E_2-E_1)$ and assuming ΔE_2 to be so small

[5] S. Chandrasekhar, Revs. Modern Phys. 15, 1 (1943).
[6] H. Lorentz, Proc. Amsterdam Acad. Sci. 8, 591 (1906).
[7] (a) W. Lenz, Z. Physik 25, 299 (1924); (b) V. Weisskopf, Z. Physik 75, 287 (1932).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

that the perturber is deflected but very little from its straight-line path, then, if it flies by at a closest distance ρ from the radiator and has a speed v,

$$\Delta\varphi=\hbar^{-1}\int\epsilon(\rho^2+v^2t^2)^{\frac{1}{2}}dt=\Delta\varphi(\rho,v). \qquad (1.5)$$

The added phase depends on ρ and v. An *effective* impact, i.e., one which kills coherence, is an impact for which $\Delta\varphi\gtrsim\varphi_0$, a quantity presumably of order of magnitude 1. Thus the relation $\Delta\varphi=\varphi_0$ divides all impacts into two ideal classes, those which broaden and those which do not. Since a smaller ρ leads in general to a larger $\Delta\varphi$, the relation $\Delta\varphi=\varphi_0$ defines a $\rho_c(v)$ *below* which all impacts broaden the line. Therefore $\pi\rho_c^2(v)$ plays the role of Lorentz' q and may be introduced in (1.3) and (1.2). Here then is a qualitative reason why optical cross sections may differ from gas kinetic ones, and why they depend on v.

We shall often encounter the equation $\Delta\varphi=\varphi_0$ as a limiting condition separating the ranges in which different kinds of treatment are necessary. The precise value of φ_0 never matters, but we assume it to be of order magnitude 1 (or $\pi/2$, or even π) and call it the *critical phase.* If $\Delta\varphi$ arises from single impacts, the critical phase defines an *optical radius* or *critical impact distance* ρ_c, which depends on v and on the form of ϵ, i.e., the law of interaction. Likewise, there results an *optical cross section* $\pi\rho_c^2$.

Numerous refinements of this theory have been proposed.[4b,8] One of them is treated later in some detail because of its success in applications even under circumstances in which its *a priori* validity is not evident (see Sec. IV).

C. Relation Between Statistical and Impact Theories

The two physical pictures underlying the treatments in parts A and B of this section are so completely unrelated that a decision as to their domains of validity cannot be made on simple intuitive grounds. Attempts to combine the results in phenomenological ways, by merely saying that both are present independently and thus "folding" one distribution into the other[2d,4b] have had some success,† but lack fundamental justification. It is necessary to fall back upon a mathematical formalism more general than either treatment A or B, and to see under what conditions it reduces to these species of description.

Such a formalism is available in the Fourier integral for the line width, suitably generalized.[9,10] A classical observable Q, which is a function of momenta and coordinates of some physical system, is represented in quantum mechanics by a matrix with elements Q_{ij}. The classical intensity distribution within a line,

$$I(\omega)=\frac{2\omega^4}{3\pi c^3}\left|\int_{-\infty}^{\infty}\mu(t)e^{-i\omega t}dt\right|^2, \qquad (1.6)$$

is such an observable involving momenta and velocities of the radiating atom through the varying dipole moment μ and the atomic frequency ω. The observed value of Q for the state of the system which is represented by the statistical matrix $\mathbf{S}$ is

$$\bar{Q}=\text{trace}(SQ).$$

Thus the quantum mechanical intensity distribution (we omit the bar) is, because of (1.6),

$$I(\omega)\propto\text{trace}\left\{S\int dt\mu(t)e^{-i\omega t}\int dt'\mu(t')e^{i\omega t'}\right\}, \qquad (1.7)$$

where S, $\mu(t)$, and $\mu(t')$ are matrices, the latter two in the so-called moving system or in "the Heisenberg form" (see the following). From (1.7), statistical and impact theories can both be derived as certain limiting cases.[8d,11]

The Hamiltonian H for the emitting atom is a function of t because of the perturbations that broaden the line. Introducing a time-development matrix U which satisfies the Schrödinger equation,

$$i\hbar\dot{U}=HU, \qquad (1.8)$$

then

$$\mu(t)=U^\dagger\mu U,$$

μ being the time-constant (Schrödinger) matrix $\mu_{ij}^0=\int\psi_i^*(\sum_n er_n)\psi_j d\tau$. Now suppose the perturbers are fixed in space. Then H does not depend on t and (1.8) has the solutions $U_{lm}=\exp[(-i/\hbar)E_l t]\delta_{lm}$, where E_l is a stationary eigenvalue of H. This approximation is too drastic and yields no line width; however, almost as simple and quite significant is the answer obtained by making the *adiabatic hypothesis.*‡

The adiabatic hypothesis in its quantum mechanical form says that, under conditions specified forthwith, Eq. (1.8) has a solution of the form

$$U_{lm}=\exp\left[-\frac{i}{\hbar}\int_0^t E_l(t')dt'\right], \qquad (1.9)$$

[8] (a) W. Lenz, Z. Physik **80**, 423 (1933); (b) C. Reinsberg, Z. Physik **111**, 95 (1938); (c) E. Lindholm, Arkiv Mat. Astron. Fysik **28B**, No. 3 (1941); (d) H. M. Foley, Phys. Rev. **69**, 616 (1946).

† This process is not the same as adding the half-widths.

[9] P. W. Anderson, Phys. Rev. **76**, 647 (1949).

[10] S. Bloom and H. Margenau, Phys. Rev. **90**, 791 (1953).

[11] (a) The procedure here follows H. Margenau and R. Meyerott, Astrophys. J. **121**, 194 (1955). See also: (b) H. Margenau, Z. Physik **86**, 523 (1933); (c) T. Holstein, Phys. Rev. **79**, 744 (1950); (d) M. Mizushima, Phys. Rev. **83**, 94 (1951); **84**, 363 (1951); (e) P. W. Anderson, Phys. Rev. **86**, 809 (1952).

‡ The word "adiabatic" has come to be almost meaningless through unprincipled use in modern physics. This becomes evident in Sec. IIB, where various specific senses of "adiabatic" and its opposite, "diabatic" (or its atrocious, pleonastic synonym, non-adiabatic), are discussed. In the present context, the meaning should be clear from the discussion.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

where the $E_l(t)$ are solutions of

$$H\psi_l(t)=E_l(t)\psi_l(t), \tag{1.10}$$

an equation sometimes called the adiabatic Schrödinger equation. The hypothesis is valid if, during the perturbation interval t, and for all states m,

$$\left|\int_0^t \frac{(dH/dt')_{lm}}{E_l(t')-E_m(t')}\left\{\exp\frac{i}{\hbar}\int_0^{t'}[E_l(t'')-E_m(t'')]dt''\right\}dt'\right|\ll 1. \tag{1.11}$$

Here l is the atomic state initially present, m some other state. With (1.9) and (1.10) $\mu(t)$ takes a simple form:

$$\left.\begin{aligned}\mu_{lm}&=\mu_{lm}{}^0\exp[i\varphi_{lm}(t)]\\ \varphi_{lm}&=\hbar^{-1}\int_0^t[E_l(t')-E_m(t')]dt'\end{aligned}\right\}. \tag{1.12}$$

We now introduce this into (1.7) and assume that S is diagonal, a form which can always be brought about by proper choice of state functions. Then, after a few steps (among them the replacement of the variables $t'-t$ by τ) there results

$$I(\omega)\propto\int d\tau e^{i\omega\tau}C(\tau), \tag{1.13a}$$

$$C(\tau)=\sum_{if}S_i|\mu_{if}{}^0|^2\int_{-\infty}^{\infty}\exp\{-i[\varphi_{if}(t+\tau)-\varphi_{if}(t)]\}dt. \tag{1.13b}$$

The function $C(\tau)$, whose Fourier transform is the intensity distribution, is called the correlation function; its general features form the substance of an elegant mathematical discipline called the statistics of time series; hence, this manner of writing $I(\omega)$ is finding increasing favor with theorists. Qualitatively, the meaning of C is clear: $\varphi_{if}(t+\tau)$ and $\varphi_{if}(t)$ are certainly equal when $\tau=0$; hence, $C(0)$ is large. If φ has a short "memory" so that it quickly loses correlation with its previous values, $C(\tau)$ will fall off rapidly with τ. It is then something like a δ function and $I(\omega)$ is broad. If, on the other hand, $\varphi(t)$ remains correlated with itself (in the absence of interaction there is a linear correlation for all time: $\varphi=\text{const}\cdot t$) during an appreciable period, $C(\tau)$ will differ from 0 for that period, and $I(\omega)$ will be correspondingly narrow. One may indeed show that the half-width of $I(\omega)$ is approximately the reciprocal of the half-width of $C(\tau)$.

In arriving at (1.13), which is suggestive of the starting point of the Lorentz theory, we have used only the adiabatic hypothesis. Its correctness is tied to inequality (1.11).

An analysis of the conditions under which that inequality is true is given in the next section. We observe here primarily that (1.11) contains both the nondiagonal element $(dH/dt)_{lm}$ and the energy difference $E_l(t)-E_m(t)$; its validity depends on the relation between them. The remainder of this section takes (1.11) for granted and concerns itself with the manner in which E_l-E_m changes in time, for the applicability of statistical and impact theories depends on the behavior of that difference alone.

The statistical theory results if the difference changes slowly. The quantity φ may then be expanded in a Taylor series

$$\varphi(t+\tau)=\varphi(t)+\varphi'(t)\cdot\tau+\cdots,$$

and all terms in higher powers of τ may be neglected. The convergence here is generally good if only $\varphi''(t)(\tau^2/2)\ll 1$ for all values of t and those values of τ which matter in the integration of (1.13a), i.e., for which $C(\tau)\neq 0$. In (1.13b) we assume for simplicity that only one atomic state $i=1$ was initially present ($S_i=\delta_{i1}$) and that only one $\mu_{1f}{}^0$, namely $\mu_{12}{}^0$, is of appreciable magnitude. Then

$$C(\tau)\propto\int e^{i\varphi'_{12}(t)\tau}dt=\int\exp\{-i[E_2(t)-E_1(t)]\tau/\hbar\}dt;$$

hence,

$$I(\omega)\propto\int d\tau e^{i\omega\tau}\int\exp\{-i[(E_2-E_1)/\hbar]\tau\}dt.$$

The integral over τ now leads to a δ function which is different from zero only at those times t when ω, the line frequency under consideration, equals $[E_2(t)-E_1(t)]/\hbar$, our former $\omega(r_1\cdots r_N)$. By the rules of statistical mechanics (ergodic theorem) these times have probabilities equal to the fractional volume of configuration space in which $\omega(r_1\cdots r_N)=\omega$. The exact formal analysis here involves ranges $d\omega$, near equalities and probability densities, which we suppress, placing thereby a heavy burden of infinities and zeros upon the proportionality signs. With a little care the reader can supply these purely mathematical details. In sum

$$I(\omega)\propto\int\cdots\int\delta[\omega,\omega(r_1\cdots r_N)]d\mathbf{r}_1\cdots d\mathbf{r}_N, \tag{1.14}$$

and this is the $P(\omega)$ of Eq. (1.1).

An extension of the reasoning here outlined permits proof of two propositions of interest in the use of the statistical method.[1f,8a,11]

1. Far in the wings of a line, i.e., as $|\omega-\omega'|\to\infty$, ($\omega'$ is again the normal frequency) the statistical theory is always applicable. The value of $|\omega-\omega'|$ beyond which this is true depends on the form of the interaction and cannot be generally specified.

Later we have occasion to return to this proposition (which we call for short the *wing theorem*) and therefore give its more detailed proof, which closely follows

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

Sobel'man. Substituting Eq. (1.13b) into (1.13a) and dropping the subscripts and the sum we have

$$I(\omega) \propto \int d\tau e^{i\omega\tau} \int e^{-i[\varphi(t+\tau)-\varphi(\tau)]} dt,$$

with the understanding, however, that this quantity must be averaged over different collisions.

Let us *redefine* φ *(without bothering to change notation) to mean only the variable part of the phase which is contributed by the perturbation.* We then have for the second integral

$$\int e^{-i[\omega'\tau+\varphi(t+\tau)-\varphi(t)]} dt.$$

On interchanging the order of integrations, we have

$$I(\omega) \propto \int e^{i\varphi(t)} dt \int d\tau e^{i[\Delta\omega\tau-\varphi(t+\tau)]},$$

where $\Delta\omega=\omega-\omega'$. We now set $t+\tau=s$ in the second integral and obtain

$$I(\omega) \propto \int e^{i(\varphi-\Delta\omega t)} dt \int e^{-i[\varphi(s)-\Delta\omega s]} ds.$$

In the wings of the line, i.e., for sufficiently large $\Delta\omega$ the contributions to

$$\int e^{i(\varphi-\Delta\omega t)} dt$$

come from intervals Δt_k around the points t_k, defined by $\varphi'(t_k)=\Delta\omega$ since in other regions the rapid oscillations of the integrand tend to cancel. We can then write

$$\int e^{i(\varphi-\Delta\omega t)} dt \doteq \sum_k \int_{\Delta t_k} e^{i[\varphi-\Delta\omega t]} dt.$$

We next expand the exponent in a Taylor series about t_k obtaining

$$\sum_k e^{i[\varphi(t_k)-\Delta\omega t_k]} \int_{\Delta t_k} \exp i\left[\frac{\varphi''(t_k)}{2!}(t-t_k)^2 + \frac{\varphi'''(t_k)}{3!}(t-t_k)^3+\cdots\right] dt.$$

The important range of the integration, Δt_k, is of the order

$$\sqrt{2}[\varphi''(t_k)]^{-\frac{1}{2}}.$$

If, in this range, the term proportional to $(t-t_k)^3$ is small, the integration can be extended to infinity. This condition is

$$\varphi'''(t_k)[\varphi''(t_k)]^{-\frac{3}{2}} \ll 1. \qquad (1.15)$$

We then obtain for the integral,

$$\sum_k e^{i[\varphi(t_k)-\Delta\omega t_k]}[\varphi''(t_k)]^{-\frac{1}{2}},$$

and, for $I(\omega)$,

$$I(\omega) \propto \sum_{kl} e^{i[\varphi(t_k)-\varphi(t_l)-\Delta\omega(t_k-t_l)]}[\varphi''(t_k)\varphi''(t_l)]^{-\frac{1}{2}}.$$

The average over all collisions causes the terms for which $k \neq l$ to cancel, and we find

$$I(\omega) \propto \sum_k [\varphi''(t_k)]^{-1}.$$

The sum here can be interpreted as proportional to the length of time during which

$$\Delta\omega = \varphi'(t_k)$$

or

$$\omega = \omega' + \varphi' \equiv \hbar^{-1}[E_i(t_k) - E_f(t_k)].$$

Thus $I(\omega)$ is proportional to the time that the atom spends in the perturbed condition with the energy separation between levels given by $\hbar\omega$. This leads again to Eq. (1.14) by arguments already presented; hence the wing theorem.

2. If a spectral line is broadened by *single impacts* of perturbers (low density) and Δt is the time during which the passing particle moves a distance equal to the impact parameter ρ, the statistical theory is applicable provided

$$\omega_{\frac{1}{2}} \cdot \Delta t \gg 1. \qquad (1.16)$$

Here $\omega_{\frac{1}{2}}$ is the actual, measured half-width. Proof of theorem 2 rests upon the assumption that for an individual impact the perturbation can be written in the form

$$E_2 - E_1 = \text{const} r^{-\sigma},$$

where r is the distance between radiator and perturber and σ a small positive integer, so that $\varphi(t)$ can be constructed for any impact, If one then employs inequality (1.15) noticing that, by virtue of (1.13a), the half-width of $C(\tau)$ is the reciprocal of the half-width of $I(\omega)$, a little algebra leads to formula (1.16).

This formula has been proved for individual impacts only, and $\omega_{\frac{1}{2}}$ is the actual half-width of the line; in the wing theorem, the value of ω beyond which the statistical theory holds bears no general relation to the half-width and differs in different situations.

Next, we demonstrate that (1.13) leads to the impact results when the phase changes $\Delta\varphi$ occur suddenly and are therefore separated in time. We use (1.13) later on to obtain impact formulas more refined than those of Lorentz or Weisskopf, following the work of Foley, Lindholm, and others. At present we retrace the step that led to (1.13) and write

$$I(\omega) \propto \left| \int \mu_{12}^0 e^{-i[\omega t-\varphi_{12}(t)]} dt \right|^2, \qquad (1.17)$$

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

assuming again that the atom has only two levels. If sudden impacts occur at times $t=T_i$, and if these impacts add amounts $\hbar a_i$ to the phase of the atom,

$$E_2-E_1=\hbar[\omega'+\sum_i a_i\delta(t,T_i)]$$

so that

$$\varphi_{12}\equiv\varphi(t)=\omega' t+A_j \quad \text{for } T_j\leqslant t<T_{j+1},$$

where $T_1<T_2<T_3\cdots$ and

$$A_j=\sum_{i=1}^{j} a_i.$$

Then the integral in (1.17) is

$$\mu_{12}{}^0\sum_j e^{iA_j}\int_{T_j}^{T_{j+1}} e^{-i(\omega-\omega')t}dt.$$

When the square of the absolute value of the sum is taken, there will appear the terms

$$|\mu_{12}{}^0|^2\sum_j\left|\int_{T_j}^{T_{j+1}} e^{-i(\omega-\omega')t}dt\right|^2 \tag{1.18}$$

and, in addition, cross terms with factors $e^{i(A_k-A_j)}$. If the a_i are unrelated, the sum of these will effectively banish. The remaining summation in (1.16) extends over different free times $T_{j+1}-T_j=T$ and is equivalent to the integration over T in Eq. (1.2). Under these assumptions the Lorentz formula is therefore established.

The wings of a line rarely follow the Lorentz formula, whereas the central core often conforms to it. The reason for this is not hard to see.

We spoke of sudden phase changes $\Delta\varphi$. In quantitative terms, "sudden" means that the change in $\varphi(t)$ occurring in the integrand of Eq. (1.17) as the result of a collision is limited to a time interval Δt small compared to the period of the other exponential factor of the integrand, i.e.,

$$\Delta t\ll 2\pi/\omega. \tag{1.19}$$

Unless the collisions overlap completely (long-range forces, high density of perturbers) this inequality is satisfied for sufficiently *small* ω, that is to say, near the center of the line. This result might be called the "core theorem" in contradistinction to the wing theorem. The former says that the core of a line "results from impacts," while the latter *claims that the wings* "are statistical."

These considerations allow these qualitative inferences: impact theories describe lines at high temperatures and for sudden perturbations, i.e., at low densities; statistical theories meet success at low temperatures and for heavy perturbers, especially at high densities where many interactions coincide and produce small net fluctuations of energy at the radiating atom.

To account for the contour of a given spectral line when neither impact nor statistical theory can be fully trusted, the center of the line can usually be interpreted as resulting from impacts, the wings as statistically broadened. More is said about this in Sec. IV.

We should comment on the consequences of Eq. (1.7) when the adiabatic hypothesis cannot be made, that is, when condition (1.11) is violated. This contingency is most likely to arise when the spacings, E_l-E_k, are small, hence in the microwave region, and under other more special circumstances. The story is then somewhat involved: the best account of it is to be found in Anderson's article.[9]

In the plasma situation the statistical theory promises a satisfactory description of the ion effects, and it would seem that impact theories might be suited to treat the swift electrons. But in this latter task, unfortunately, a fundamentally new problem arises. All theories here surveyed require the use of a potential energy of interaction ϵ, which is a function of the perturber position. Basically, then, they assume that the perturber has a classical path. This is not true in general for electrons, whose position is quite diffuse when the speed is definite because of the uncertainty principle. To operate with an $\epsilon(r)$ may therefore be meaningless, and a new approach is needed. In the next section, we offer qualitative arguments designed to permit some discrimination of conditions under which classical path theories are useful, and related considerations.

II. CLASSICAL-PATH HYPOTHESIS AND CRITERIA FOR ADIABATICITY

A. Classical Path

Essential to the foregoing analysis of spectral line broadening are two assumptions: first, that the perturber has a classical path, i.e., is a point particle with spatial coordinates that are functions of the time; and, secondly, that adiabaticity prevails, so that the perturber disturbs (without mixing) only the two states between which the system is radiating. The lack of a completely general workable approach to the broadening problem confers a good deal of importance upon a knowledge of the conditions under which such assumptions hold. As instances of treatments where these assumptions are not invoked we note the quantum theories of Jablonski[12] and of Kivel, Bloom, and Margenau[13]; another example is the theory of dielectric relaxation of Debye,[14] which treats the broadening induced by reorientation transitions in the radiator during collisions. The broadening mechanisms pertaining to these latter theories are quite different from those discussed in Sec. I. The two classes of results arise from opposite extremes in the physical model for the

[12] A. Jablonski, Phys. Rev. 68, 78 (1945).
[13] Kivel, Bloom, and Margenau, Phys. Rev. 98, 495 (1955).
[14] (a) P. Debye, *Polar Molecules* (Chemical Catalog Company, Inc., New York, 1929), Chap. V; (b) J. H. Van Vleck and V. Weisskopf, Revs. Modern Phys. 17, 227 (1945).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

interaction. We now formulate some simple criteria for the validity of the description of Sec. I. Special consideration is given to the interaction of a radiating hydrogen atom with electrons and protons.

Quantum mechanics yields an accurate account of atomic phenomena, and, through the uncertainty principle, also defines the limits of classical mechanics. This principle focuses attention on our inability to describe certain details of an event, such as the collision between two systems, without introducing finite disturbances in the parameters necessary for such description. When these disturbances compete in magnitude with the events one wishes to investigate, the event imagined cannot be observed in the degree of detail demanded, for then a classical picturization is without basis, and we must use quantum mechanics.

Classical-path theories assume that the distance of separation r between the perturber and the radiator is expressible as

$$r(t)=(\rho^2+v^2t^2)^{\frac{1}{2}}, \tag{2.1}$$

which requires exact specification of both its position and velocity at each instant of time. For this is to be meaningful the uncertainties in these quantities, as derived from quantum considerations, must be small in comparison with their actual magnitudes.

Numerous authors base their use of (2.1) on the magnitude of the angular momentum involved in a collision,[1c,8d,15] relying on the correspondence principle to validate classical analysis for large orbits. If $\mathbf{L}$ is the angular momentum of the perturber with respect to a radiator at rest, then, $L\approx l\hbar$. For large values of the integer l we may identify $\mathbf{L}$ with $m\mathbf{r}\times\mathbf{v}$, which during the collision may be taken to be $\approx mrv$.

The relation above in the form

$$mrv\approx l\hbar$$

and the uncertainty principle,

$$m\Delta r\Delta v=\hbar,$$

yield on division

$$(\Delta r/r)\cdot(\Delta v/v)=1/l, \tag{2.2}$$

and this is much smaller than 1 if l is large.

Equation (2.2) implies that for sufficiently large l any specified relative accuracy in r and v can be accomplished without violating the uncertainty principle, i.e., without need for abandoning the assumption of a classical path. For example, if $l\approx 100$ we may choose the orbit of the perturber such that $\Delta r/r\approx\Delta v/v\approx 1/10$, which ratios represent an acceptable percentage accuracy in the specification of these parameters.[9]

The preceding argument leaves many detailed questions unanswered. Let us consider the problem with a measure of skepticism which, fortunately, will later prove excessive. For the process of collision broadening to be tractable by means of a classical path we must require that the inequality

$$(\Delta r/r)\cdot(\Delta v/v)\ll 1 \tag{2.3}$$

holds throughout the time of any collision and for all impact radii of importance. When the classical path is derived properly as the limiting form of a quantum mechanical description these requirements lead to relations more specific than (2.3), relations which depend upon the mass of the perturbers, the density and temperature of the gas, and the type of interaction which pertains to the problem.

Collisions with large angular momenta, as shown by Eq. (2.2), can be described classically; this means that the perturbing particles correspond to wave packets, well localized both in coordinates and momentum, the limits of this localization being Δr and Δv. We now imagine a wave packet with a mean spatial extension a and a mean momentum mv, this latter quantity being equal to the actual momentum of the classical particle. In addition, we let r be the distance between the radiator and the center of the packet; it therefore corresponds to the collision distance previously denoted by the same symbol. For the validity of the classical description we demand that

$$a\leqslant r, \tag{2.4}$$

the ultimate limit of the classical description being $a\approx r$.

In all applications of the theory the interaction is taken to be vanishingly small outside of a finite region of space, the linear dimension of which is designated by d. Physically, this distance represents the extent of penetration of the field of the radiator into its surroundings. For the case of a neutral plasma this distance is the Debye length[16]

$$d=6.90(T/n)^{\frac{1}{2}}, \tag{2.5}$$

where d is given in centimeters, T in degree absolute, and n in cm^{-3}.

We consider the interaction region to be a cube of dimensions d, oriented for convenience with an axis parallel to the perturber's path. The time during which the perturbation is in effect, τ_p, is then of the order of d/v, a constant for all values of r up to approximately d, beyond which distance it vanishes.

In a classical description of the passage of the perturber through the interaction region, the collision time must be shorter than the time during which the perturber diffuses quantum mechanically through the distance r. This diffusion results from the initial localization of the packet and proceeds with a velocity of the order of $\hbar/ma$. It is therefore necessary that

$$r/(\hbar/ma)\gg d/v. \tag{2.6}$$

Because of (2.4) this inequality must hold for, $a\approx r$;

[15] (a) L. Spitzer, Phys. Rev. 55, 699 (1939); (b) E. Lindholm, Arkiv Mat. Astron. Fysik 32A, No. 17 (1945).

[16] L. Spitzer, *The Physics of Fully Ionized Gases* (Interscience Publishers, Inc., New York, 1956).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

hence, we get from (2.6)

$$a \gg (d\lambda)^{\frac{1}{2}}, \tag{2.7}$$

where λ is the mean de Broglie wavlength of the perturber, $\hbar/mv$.

Applying this basic inequality to the most distant collisions with least momentum uncertainty for which $r \approx d \approx a$, we obtain $a \gg \lambda$, which (since $a \approx r$) is nothing more than the condition that $l \gg 1$. Thus, the criterion (2.7) reduces to the conditions that the angular momentum be large when applied to the most distant collisions; in general, (2.7) is more restrictive than $l \gg 1$.

In addition to the condition that the passage of the perturber through the interaction region retain its meaning in view of the diffusion of wave packets, we might also require that the interaction itself be classical in terms of precise momentum exchanges.

The interaction energy between the perturber and the radiator is $V(r)$ in the classical description. This can usually be expressed in the form

$$V(r) = C_\sigma / r^\sigma,$$

C_σ being some constant. The momentum exchange for the collision is then

$$\Delta p = -\int \nabla V(t) dt.$$

Another quantity of interest in this connection is the phase change per collision

$$\Delta\varphi = \hbar^{-1} \int V(t) dt.$$

In both these integrals we replace dt by dx/v and $V(t)$ by $C_\sigma(\rho^2 + v^2t^2)^{-\sigma/2}$. The integrations should then extend from $-d/2$ to $d/2$, but a sufficient approximation results when they are performed from $-\infty$ to $+\infty$. Curvature of the path is neglected in this procedure. Elementary calculation yields

$$\Delta p = \frac{\sigma C_\sigma}{v\rho^\sigma} \pi^{\frac{1}{2}} \frac{\Gamma[(\sigma+1)/2]}{\Gamma[(\sigma+2)/2]} \approx \frac{V(\rho)}{v}$$

and

$$\Delta\varphi = \frac{C_\sigma}{\hbar v \rho^{\sigma-1}} \pi^{\frac{1}{2}} \frac{\Gamma[(\sigma-1)/2]}{\Gamma(\sigma/2)}.$$

The critical impact distance ρ_c is that value of ρ for which $\Delta\varphi = 1$ (see the end of Sec. IB). Solving this equation for C_σ in terms of ρ_c and inserting in Δp, we find

$$\Delta p = (\rho_c/\rho)^{\sigma-1}(\sigma - 1/\rho)\hbar. \tag{2.8}$$

To regard the perturber wave as a classical particle is to neglect its momentum uncertainty $\hbar/a$. The same uncertainty is thus introduced also in the states of the radiator system during and after the collision. The validity of the classical interaction is certainly assured if we require that all momentum (and ther corresponding energy) exchanges between the pe and the radiator exceed this neglected uncertain classical picture therefore holds only when:

$$\Delta p \approx V(\rho)/v \gg \hbar/a.$$

This places an *upper* limit on the magnitude o which the interaction may be treated classically.

When the opposite inequality

$$\Delta p \approx V(\rho)/v \ll \hbar/a$$

holds, the momentum uncertainty in the syste exceeds the momentum exchanges of the cl description, and the spread in momenta must be porated into the description of the perturber (or, is equivalent, the spatial spread of the packet m taken into account). Then the classical picture collision breaks down.

From (2.9a) using the more accurate expressio Δp, from Eq. (2.8), we have

$$\rho \ll \alpha_\sigma \rho_c (a/\rho)^{1/(\sigma-1)},$$

where $\alpha_\sigma = (\sigma-1)^{1/(\sigma-1)}$ is of the order of unity, a in the correlation of σ with α_σ below:

$\sigma =$	2	3	4	6
$\alpha_\sigma =$	1.00	1.41	1.44	1.38

Since $a \leqslant r$ for all r during the collision, and r can on the value ρ, (2.10) seems to imply

$$\rho \ll \alpha_\sigma \rho_c. \quad (2.11)$$

This is essentially the condition $\Delta\varphi \gg 1$. We con that only collisions within the optical radius amenable to a classical-path treatment and all which occur outside of ρ_c must be treated by mea quantum mechanics. We return to this somewhat prising conclusion at the end of this section.

Letting $a \approx \rho$ as the least stringent condition fo classical path we obtain in place of (2.10)

$$a \ll \alpha_\sigma \rho_c. \quad (2.12)$$

The validity of the classical description for colli within ρ_c is also contingent on the enforcement of On combining (2.7) and (2.12), there results on nating,

$$(d\lambda)^{\frac{1}{2}} \ll \rho_c. \quad (2.13)$$

This condition for the validity of the classical pa effectively independent of the actual size of the packet describing the perturber. Margenau and Ki actually computed the dimension of a one-dimens packet [see Eq. (2.23)] and required it to be less the size of the radiator. In this way they found the width of the Lyman-α line of hydrogen ma computed classically at $T \approx 6000°$K if the pertu mass $> 10^{-26}$ g, which is true for protons but no

[17] H. Margenau and R. Kivel, Phys. Rev. 98, 1822 (1955

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by eGangotri Foundation USA

TABLE I. Broadening of H_α by first-order Stark effect. Temperatures for protons and electrons *below* which the classical-path treatment holds, for various values of the density.

n (cm^{-3})	(°K) (protons)	(°K) (electrons)
10^{12}	4×10^6	50
10^{14}	4×10^7	500
10^{16}	4×10^8	5000
10^{18}	4×10^9	50 000

TABLE II. Broadening by second-order Stark effect. Temperatures for protons and electrons below which the classical-path treatment holds, for various values of the electron (ion) density.

n (cm^{-3})	(°K) (protons)	(°K) (electrons)
10^{12}	10^7	0.1
10^{14}	10^{10}	100
10^{16}	10^{15}	10^5
10^{18}	10^{16}	10^8

electrons. All this is in keeping with the present result, which is derived in a more general way.

After substituting (2.5) along with the definitions of λ and ρ_c into (2.13) one finds

$$\frac{T^{1/(\sigma-1)}}{n^{\frac{1}{2}}M^{\frac{1}{2}+1/(\sigma-1)}} \ll \frac{[\gamma_\sigma(C_\sigma/\hbar)]^{2/(\sigma-1)}}{6.90\hbar(3k)^{-\frac{1}{2}+1/(\sigma-1)}}, \tag{2.14}$$

where v, the perturbers' velocity, has been taken to be its rms value

$$v=(3kT/M)^{\frac{1}{2}},$$

M is the mass of the perturber, either a proton or an electron, and the γ_σ are numbers of order unity (e.g., $\gamma_2=\pi$).

This result has the advantage of containing no reference to ρ, which has no experimental meaning. However, inequality (2.14) is a sufficient condition which is in some instances too severe. For hydrogen, the first order Stark effect ($\sigma=2$) involves[18]

$$C_2\approx\tfrac{3}{4}(\hbar^2/m)n'(n'-1),$$

where n' is the principal quantum number of the excited state of the radiating atom and m is the electron mass. This is a sufficient approximation. Accurately, $n'(n'-1)$ should be replaced by $N_{n'}$ such that

$$N_3=4.5,\quad N_4=9.9,\quad N_5=23.6,\quad N_6=31.8.$$

Hence (2.14) becomes

$$\frac{T}{n^{\frac{1}{2}}M^{\frac{1}{2}}} \ll \frac{(\pi C_2)^2}{6.90\hbar^3(3k)^{\frac{1}{2}}}.$$

For the first Balmer line (H_α, $n'=3$), the inequality (2.14) gives Table I. For the higher Balmer lines the temperatures appearing here are multiplied by the factor $[n'(n'-1)/6]^2$, $n'>3$. Hence, for the higher series members the classical-path treatment is valid over a greater temperature range. This increase has practical significance only for the electrons, since they alone are severely limited by the values of the temperature. For a pure hydrogen plasma, thermal ionization begins at about 7000°K, so that appreciable broadening by electrons and protons does not occur at temperatures below this value. Thus much of Table I is purely academic.

For a typical second-order Stark effect in the optical region ($\sigma=4$) $C_4\hbar^{-1}\approx2\times10^{-12}$ esu. Here we obtain from (2.14) $T^{1/3}\ll n^{1/2}M^{5/6}(\gamma_4C_4/\hbar)^{2/3}(3k)^{1/6}/6.9\hbar$ and $\gamma_4=\pi/2$. This leads to the values of n and T shown in Table II. Table II shows that electrons may be treated classically for $n>10^{15}$ cm^{-3} in the second-order Stark effect, while protons and heavier ions always permit such treatment.

B. Criteria for Adiabaticity

We now transfer attention to the radiating system. Perturbation of its states can be treated adiabatically when the inequality (1.11) holds. The state of the system is then maintained throughout a collision except for its alteration through the radiation process. There are two cases of interest: (1) the states of the unperturbed radiator may be classified into groups of degenerate states and (2) the states of the unperturbed radiator are all nondegenerate. In case 1, two questions can be asked: (1a) Does the perturbation cause transitions between two states containing within the same group of degenerate levels? (1b) Does the perturbation cause transitions between a degenerate state in one group and a degenerate state in another group separated from the first by a finite energy? The perturbation may well be adiabatic in the sense of 1b, diabatic in the sense of 1a. In this section, we treat only case 1a, leaving 1b, which is more difficult, for the most part until Sec. V. Only a brief qualitative remark concerning 1b is presented at the end of this part of Sec. II. Existence is assumed of a finite perturbation time τ_p, the rate of change of the perturbation dV/dt being different from zero only during that time which, as we have seen, is of the order d/v. In addition, we here treat single impacts only.

For case 1a the energy differences of the instantaneous states appearing in the condition for adiabaticity are due solely to the perturbation itself, since all other states of the radiator may be ignored because of their much larger energy separations.

As the simplest example, we consider a radiator whose excited initial state has a two-fold degeneracy, and this degeneracy is removed in the first order of the per-

[18] A. Unsöld, *Physik der Sternatmosphären* (Springer-Verlag, Berlin, Germany, 1955), second edition, p. 321.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

turbation theory. The adiabatic condition then becomes

$$\left|\int_{-\infty}^{+\infty}\frac{d/dt'\,|V_{12}(t')|}{2|V_{12}(t')|}\times\exp\left[-\frac{i}{\hbar}\int_{-\infty}^{t'}2|V_{12}(t'')|\,dt''\right]dt'\right|\ll 1, \quad (2.15)$$

where $V_{12}(t)=\langle 1|V(t)|2\rangle$, the indices 1 and 2 designating the two states in question.

The phase shift between the states is given by

$$\Delta\varphi(t)=\frac{2}{\hbar}\int_{-\infty}^{t}|V_{12}(t')|\,dt'.$$

When $|\Delta\varphi(t)|\gg 1$, there have been many oscillations of the exponential part of the integrand in (2.15) at time t. These oscillations cause the integral to be quite small provided that

(α) $(d/dt)|V_{12}(t)|/|V_{12}(t)|$ is a slowly varying function of the time during an average period of the oscillation $\hbar/2\langle|V_{12}|\rangle$ (where $\langle|V_{12}|\rangle$ is a time average of $|V_{12}|$ for the whole perturbation,

$$\langle|V_{12}|\rangle=\int_{-\tau_p/2}^{\tau_p/2}|V_{12}(t)|\,dt/\tau_p);$$

and

(β) the perturbation time is much greater than $\hbar/2\langle|V_{12}|\rangle$.

The first requirement α reduces to the second β when the classical-path procedure is used. In that case

$$V_{12}(t)\equiv C_{12}/(\rho^2+v^2t^2)^{\sigma/2}, \quad \tau_p\leqslant t\leqslant\tau_p/2$$
$$\equiv 0, \quad |t|>\tau_p/2$$

and $(d/dt)|V_{12}|/|V_{12}|=(-\sigma v^2t)/(\rho^2+v^2t^2)$. We wish this to vary slowly within an interval of the order of $\hbar/\langle|V_{12}|\rangle$. This can occur only when $|t|\gg\hbar/\langle|V_{12}|\rangle$. To maintain condition α, therefore, the contribution of $(d/dt)\ln|V_{12}|$ coming from small $|t|$ must be inappreciable, and this requires that $\tau_p\gg\hbar/\langle|V_{12}|\rangle$, which is condition β. The two conditions are not independent.

When the degree of degeneracy is greater than 2, E_l-E_m is no longer to be identified with $2|V_{lm}|$, but it is still of the same order of magnitude. The result just obtained is therefore quite general in degenerate situations, and we must depend entirely on long, slow collision processes for the validity of the adiabatic assumption. Our criterion for adiabaticity in case 1a, when written in a more general way, is, therefore,

$$\hbar^{-1}\langle|E_l-E_m|\rangle\tau_p\gg 1. \quad (2.16)$$

But $\hbar^{-1}\langle(E_l-E_m)\rangle$ is the average frequency shift of the initial state during the collision. Neglecting the frequency shift in the lower state of the radiative transition, condition (2.16) takes the form

$$\Delta\varphi\gg 1. \quad (2.17)$$

Hence, in the presence of degeneracy, collisions which fall within the optical radius are adiabatic, others are diabatic. This result has been derived by Spitzer[19] in more detailed calculations for the L_α line.

In the absence of degeneracy, case 2, the energy differences (E_l-E_m) are much larger than those which were considered in case 1. We may thus take $E_l-E_m\approx E_l^0-E_m^0$, where E_l^0 and E_m^0 are the unperturbed states of the radiator. With this understanding, inequality (1.11) reads

$$\frac{1}{|E_l^0-E_m^0|}\left|\int_{-\infty}^{+\infty}\left(\frac{d}{dt'}V_{lm}(t')\right)\times\exp\left[-\frac{i}{\hbar}(E_l^0-E_m^0)t'\right]dt'\right|\ll 1. \quad (2.18)$$

Here it is convenient to distinguish between slow and fast collisions with respect to the oscillatory period $\hbar/|E_l^0-E_m^0|$. For slow collisions, $\tau_p\gg\hbar/|E_l^0-E_m^0|$.

Integrating (2.18) by parts and making use of the fact that $V(\pm\infty)=0$ we have

$$\frac{1}{\hbar}\left|\int_{-\infty}^{+\infty}V_{lm}(t')\exp\left[-\frac{i}{\hbar}(E_l^0-E_m^0)t'\right]dt'\right|\ll 1. \quad (2.19)$$

Therefore, when $V_{lm}(t)$ varies insignificantly over the period $\hbar/|E_l^0-E_m^0|$ the inequality (2.19) will certainly be true provided $\tau_p\gg\hbar/|E_l^0-E_m^0|$, which is our premise. This assumption about V_{lm} holds for times such that $|t|>\hbar/|E_l^0-E_m^0|$ and that the adiabatic criterion reduces merely to $\tau_p\gg\hbar/|E_l^0-E_m^0|$. Hence, all slow collisions are adiabatic.

For fast collisions, $\tau_p\ll\hbar/|E_l^0-E_m^0|$, and we may take the exponential appearing in (2.18) equal to unity. This yields, upon integration of (2.18),

$$|V_{lm}(\infty)-V_{lm}(-\infty)|/|E_l^0-E_m^0|<1,$$

and the quantity on the left is always zero. The adiabatic criterion is therefore satisfied also when the perturbation time is short, provided the radiator has large energy separations.

Thus both very slow and very fast collisions are adiabatic when degeneracy (or near degeneracy) is absent in the levels concerned. In the case of degeneracy only the slow collisions are adiabatic. In this latter instance the condition for adiabaticity is similar to one of the conditions for the classical-path approximation, namely $\Delta\varphi\gg 1$.

Under the subject of degeneracy, case 1a, special attention must be given to *space* quantization, the fact that different orientations of the total angular mo-

[19] L. Spitzer, Phys. Rev. **58**, 348 (1940).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

mentum vector correspond to the same energy of the radiating system. The meaning of the term adiabatic is a little confusing in this case, as the following consideration shows.

Suppose the atom is radiating in a $^3P_1$ state, and assume first that an external magnetic field is present. The orientations in space for which $m=-1, 0, 1$ will then have slightly different energies. The passage of a perturber past the atom is said to be adiabatic if the interaction with the perturber does not transfer the radiator from one m value to another or, classically speaking, if the radiator's orientation remains fixed in space during a perturber passage. If we retain this meaning of the word adiabatic even in the absence of an external field, then an adiabatic perturbation is one which does not succeed in turning the radiator about, or in altering the value of m.

The interaction between the radiator and the perturber depends on the value of a quantum number m. However, this m is not identical with the foregoing, which refers to some fixed direction: the interaction depends on space quantization with respect to the line joining radiator and perturber, and this line moves in space as the latter flies by, turning through an angle of 180° during a complete passage. We call the space quantum number relative to this *moving* line, m'. Adiabaticity with respect to m' is clearly different from the former version because it implies rotation of the radiating atom. Spitzer,[19] Lindholm,[20] and Unsöld[18] use the term in the latter sense.

For the present problem this choice is more appropriate because in the derivation of our criterion we have taken $|E_1-E_2|$ to be $|V_{12}|$, and this implies that the radiator is in a state for which the linear combination of m eigenfunctions changes in time while m' remains fixed. With this understanding, inequalities (2.15) and (2.16) tell us under what conditions m' is unaltered by an impact.

But if m' is not changed, so that the impact is adiabatic in the second sense, then, since the radiation has to be analyzed in a fixed coordinate system (the spectrograph does not revolve about the radiating atom), an adiabatic impact corresponds to a phase change of magnitude π. This seems to imply that simple impact theories, like those of Lorentz and Weisskopf, which ignore phase changes resulting from this rotation but consider phase changes of magnitude π produced by dynamic effects as singularly important, contain grave errors. Distant impacts, for example, are not counted in these impact theories; yet they should produce a phase change π if they are adiabatic in the second sense. Physical intuition argues that this cannot be. The resolution of this apparent dilemma is simple: criterion (2.16) shows clearly that distant impacts are *not* adiabatic in the second sense. This means they do not succeed in swinging the radiator around. They *are* adiabatic in the first sense.

This problem, which is practically of minor importance but presents points of fundamental interest, is considered carefully by Spitzer.[19] He shows how the impact theories manage to obtain a reasonable result by compensating errors, the errors being an unwarranted claim of adiabaticity and a neglect of the rotation effect. Further light is shed on the "rotation problem" by an analysis[21] having an entirely different aim, but which indicates nonetheless how and why the Lorentz theory takes correct account of the change in orientation of Debye dipoles induced by collisions.

Before concluding the analysis of adiabatic conditions we take a brief look at case 1b and inquire how one should deal with transitions between one set of degenerate states and another set, separated from the first by a finite energy interval. Suppose that for transitions between any two states of the initial set and also the final set,

$$\Delta\varphi \ll 1.$$

Then the perturbation is diabatic with respect to the (perturbed) degenerate states. In that case a consideration of Unsöld[1c] makes it appear plausible that an optical transition between the two separated level sets can be handled as though each were nondegenerate. The reason is that within the limit of the uncertainty principle the degenerate levels do form a single level even after they have been split by the perturbation.

To see this we remember that the adiabatic hypothesis for transitions between the originally degenerate states of one energy level breaks down when the reverse of (2.16) is true, *viz.*

$$|E_l-E_m|\,\tau_p\hbar^{-1} \ll 1.$$

This implies that the energy uncertainty in the system, of the order of $\hbar/\tau_p$, exceeds the discrete energy shifts due to the instantaneous perturbation $|E_l-E_m|$. The states which are split by the field are therefore not revealed individually through radiation processes involving them since they overlap in energy by large amounts and form what amounts to a single diffuse energy band.

The use of the impact theory when coupled with the suggestion of Unsöld is roughly equivalent to a procedure which takes account of the diffuseness in energy of these states by solving the time-dependent perturbation equations for the probabilities of states formed through quantum transitions that are caused by the collisions. Foley[8d] and Kolb[22] have done this using the classical-path assumption, while Kivel, Bloom, and Margenau,[13] and also Landwehr[23] have carried through the calculation without this assumption.

[20] E. Lindholm, dissertation, Upsala, 1942.

[21] J. H. Van Vleck and H. Margenau, Phys. Rev. 76, 1211 (1949).

[22] A. Kolb, Thesis, University of Michigan, 1957.

[23] G. Landwehr, Thesis, Yale University, 1956 (unpublished).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

In conclusion we call attention to an early calculation by D. Blochinzew[24] who investigated the transition from diabatic to adiabatic behavior of an atom perturbed by a harmonic electric field, $E_0 \cos\omega t$.

C. Construction of Wave Packets

Since the classical-path assumption (Sec. IIA) is not always easy to justify for electron-broadening applications, we discuss this problem in quantum terms.

How can the state function of the perturbing electron be constructed? Inequality (2.9b) is the necessary condition for success of the Born approximation, which is so useful in scattering theory. But (2.9b) is normally satisfied for the plasma electrons, and as (2.9a) leads to (2.10), so (2.9b) leads to its contrary which can roughly be written $\rho \gg \rho_c$. Therefore, when densities are very low and encounters as close as the critical radius are rare, conditions under which one ordinarily employs a plane wave in describing the perturbing electron seem to be satisfied. Nevertheless, the plane wave description, which has received extensive use in problems where collimated beams of particles are directed at scattering centers, is inappropriate to our problem for several reasons.

First, we are here concerned with a single electron encountering the radiator, or, as the density is increased sufficiently, with a number of electrons statistically distributed in momenta and acting simultaneously on the radiator. The single electron hardly acts like a plane wave because the latter is stationary throughout the radiation time, and where in reality the electron has a spectrum of momenta the plane wave possesses a single value. Furthermore, the plane wave does not yield any momentary asymmetry in the charge distribution about the radiator, a situation which practically eliminates the possibility of a Stark effect.[11a] As a description of the many-electron perturbation the plane wave has similar disadvantages for the treatment of broadening, except that in this case the assumption of stationarity of the perturbation is more acceptable. At large densities one would expect the fluctuations in the perturbation to be small and the plane wave picture to become more accurate.

In stellar atmospheres where $n \approx 10^{13}$/cc, the critical radius is only about 10^{-4} times the Debye cutoff at temperatures between 5×10^3 and 10^4 °K. Thus we might expect something like single impacts by the electrons, and strong objections to the use of a plane wave for the electron must arise. For such cases Margenau and Kivel[17] employed a wave packet state function. For simplicity, they construct wave packets in one dimension and determine the width of the packet from the condition of thermal equilibrium.

The quantum-statistical density matrix for the system in a volume V and for states labelled by wave numbers κ and κ', is taken to be

$$S_{\kappa,\kappa'} = P_\kappa \delta_{\kappa,\kappa'},$$
$$P_\kappa = \frac{(2\pi mkT)^{\frac{1}{2}}}{V^{\frac{1}{3}}} \exp\left[-\frac{\hbar^2\kappa^2}{2mkT}\right], \tag{2.20}$$

and the wave packet is defined by

$$\psi(x,t) = \sum_\kappa C(\kappa,t) e^{i\kappa x}. \tag{2.21}$$

Correspondence is established between the description (2.20), which is a stationary mixed case, and (2.21), which is a nonstationary pure case, by equating expectation values for the momenta

$$P_\kappa = |C(\kappa,t)|^2, \tag{2.22}$$

This procedure yields the amplitudes but not the phases of the coefficients $C(\kappa,t)$. These are chosen on the basis of a physical argument. If the collision occurs at the time $t=0$, then because at this instant the electron involved in collision has a maximum energy uncertainty, the packet should have its minimum concentration about the radiator at that time. But minimum concentration means equality of all phases, i.e., $P_\kappa = C(\kappa,0)^2$. From (2.22) and (2.21) one then finds (upon carrying out the summation over κ as an integration)

$$\psi(x,0) \leq \exp[-mkTx^2/\hbar^2], \tag{2.23}$$

which represents a Gaussian probability packet whose width is equal to $\hbar/(2mkT)^{\frac{1}{2}}$. This is precisely the mean momentum of the ensemble described by (2.20), and the *diffusion* velocity of this packet is, satisfying enough, the *root-mean-square velocity* of the ensemble. The action of such an electron is to squeeze itself about the atom at the instant of collision and then to diffuse away again after the impact time. More general wave packets of this type, centered about points distant from the radiator and having finite mean momenta, so that they move bodily and diffuse, are even better representations of the electron which may be of assistance in broadening calculations. A paper by the present authors[25] shows how the classical formula for Stark broadening arises out of a wave packet treatment as the wave packets become more and more concentrated.

D. Summary and Appraisal

Section II has dealt with rather basic and intricate matters, and has led to some slightly surprising conclusions which, if true, restrict the validity of classical approaches more severely than might have been anticipated. It may also seem amazing that inequalities of the form $\Delta\varphi > 1$ and $\Delta\varphi < 1$ decide so many different issues. The first defines the range of validity of the classical-path description as well as the domain of adiabaticity (in the case of degeneracy). We also meet

[24] D. Blochinzew, Physik. Z. Sowjetunion 4, 501 (1933).

[25] H. Margenau and M. B. Lewis, Phys. Rev. **106**, 244 (1957).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

it again in Sec. IV, as a criterion for the correctness of statistical theories.

The typical structure of the arguments used in this chapter is the following. We carefully specified conditions under which certain conclusion (e.g., "The classical path is a valid assumption") are true. Then we said, quite without logical justification, that when the conditions are violated the conclusion will not be true. All we should have said is that we did not prove it to be true. We did not investigate the conclusions of more general conditions, and have, therefore, no assurance that more general premises will not lead to the interdicted conclusions. Hence all inequalities here encountered are permissive, are sufficient conditions, and mean to say that, under their terms, the type of description in question is certainly adequate. They do not strictly rule out the possibility that the description may be appropriate even when the inequality fails. The logic outlined is not peculiar to our treatment but characterizes most of the arguments found in the literature. They provide guides, but no final solutions.

Fortunately, the criterion derived for classical-path description is at times too limited. But there is no way of discovering this, or of assuring it, short of proceeding with more general premises. This means actually starting with quantum theory in the initial phases of the calculation to show that such treatment does, in fact, yield the same results as the classical-path description.

III. HOLTSMARK THEORY

A. Review of Conventional Treatment

The oldest and, within its limits, most successful theory of line broadening by the particles of a plasma is Holtsmark's; it treats the diffusion of intensities within a line as if each part of the line arose from a Stark effect caused by the electric field associated with a temporary configuration of the moving ions. Most of the older review articles and newer texts deal with this theory in considerable detail; we, therefore, merely sketch its features and develop as much as is needed for present purposes.

A simple approximation gives the frequency distribution at large distances from the line center extremely well. This takes into consideration only close binary encounters between the radiating atom and one ion, leaving out of account the numerous—and therefore highly probable—collaborative but weaker perturbations of many ions further away. Thus it falsifies the center but describes adequately the wings of the line. We call this approximation the *binary* form of the Holtsmark theory and present it as follows.

Selecting the radiating atom as center, we describe a sphere of radius r about it. Denote by $P(r)$ the probability that there is at least one ion within it. To obtain $P(r)$ we first focus attention on its complement $P_-(r)$, namely, the probability that there shall be *no* particles at all within r. Clearly, $P(r)=1-P_-(r)$. By the laws of combination of probabilities,

$$P_-(r+dr)=P_-(r)\cdot p_-, \tag{3.1}$$

where p_- is the probability that there be no ion between r and $r+dr$. But we know p_-: it is 1 less the probability for the presence of particles, one or more, within the shell dr. Hence, if n is the number density of ions,

$$p_-=1-4\pi nr^2dr-(4\pi nr^2dr)^2-\cdots,$$

and the powers of dr beyond the first, which represent the probabilities for the occurrence of 2, $3\cdots$ particles in dr, can be omitted for sufficiently small dr. Thus Eq. (3.1) takes the form

$$P_-(r)+P_-'(r)dr=P_-(r)[1-4\pi nr^2dr],$$

whence on integration $P_-(r)=C\exp[-(4\pi/3)nr^3]$. The constant C must be adjusted so that $P_-(0)=1$ and is therefore 1. Hence $P(r)=1-\exp[-(4\pi/3)nr^3]dr$ and the probability that the shell contain at least one ion is

$$dP(r)=4\pi nr^2\exp[-(4\pi/3)nr^3]dr. \tag{3.2}$$

It is also apparent that this is the *nearest* ion, for the factor $4\pi nr^2dr$ expresses the indiscriminate likelihood of the presence of an ion between r and $r+dr$, while the factor $\exp[-(4\pi/3)nr^3]$ conjoins it with the condition that there be *no* ion *within* r. The latter factor is often insignificant numerically and is sometimes omitted; however, it is necessary if P is to be correctly normalized. Equation (3.2) leads to $\int_0^R dP(r)=1-e^{-N}\to 1$ if R is the volume containing $N=n(4\pi/3)R^3$ particles, whereas $\int_0^R 4\pi nr^2dr=N$.

Introducing the abbreviation

$$(4\pi/3)n=r_0^{-3}. \tag{3.3}$$

Equation (3.2) may be written

$$dP(r)=\exp[-(r/r_0)^3]d(r/r_0)^3. \tag{3.4}$$

The distance r_0 thus defined is the radius of a sphere whose volume equals the mean volume per ion; it is a little smaller than the mean distance between ions, $n^{-\frac{1}{3}}$.

Equation (3.4) is the probability distribution in r. Since a given r defines an electric field, F, and an electric field defines a frequency displacement via the Stark effect, that equation also represents the distribution of the last two quantities. F is given by $F=Ze/r^2$; so substitution $r^2=Ze/F$ in (3.4) converts it to a probability distribution in F

$$\begin{aligned} dP(F)&=\exp[-(F_0/F)^{\frac{3}{2}}]d(F_0/F)^{\frac{3}{2}}\\ &=(3/2F)(F_0/F)^{\frac{3}{2}}\exp[-(F_0/F)^{\frac{3}{2}}]dF. \end{aligned} \tag{3.5}$$

Here $F_0=Ze/r_0^2$, and the minus sign was inserted in order that the normalization be correct relative to F, not F_0/F. That is to say, the normalization is such that

$$\int_0^{F=\infty} dP(F)=1.$$

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

In the *linear* Stark effect,

$$\Delta\omega = sF,$$

s being a constant. Writing $\Delta\omega_0 = sF_0$, Eq. (3.5) leads to

$$dP(\Delta\omega) = -\exp\left[-\left(\frac{\Delta\omega_0}{\Delta\omega}\right)^{\frac{3}{2}}\right] d\left(\frac{\Delta\omega_0}{\Delta\omega}\right)^{\frac{3}{2}}$$

$$= \frac{3}{2\Delta\omega}\left(\frac{\Delta\omega_0}{\Delta\omega}\right)^{\frac{3}{2}} \exp\left[-\left(\frac{\Delta\omega_0}{\Delta\omega}\right)^{\frac{3}{2}}\right] d(\Delta\omega). \quad (3.6)$$

By the fundamental assumption of the statistical theory $dP(\Delta\omega)$ is identical with the intensity distribution:

$$dP(\Delta\omega) = I(\Delta\omega)d(\Delta\omega). \quad (3.7)$$

Hence,

$$I(\Delta\omega) = \frac{3}{2\Delta\omega}\left(\frac{\Delta\omega_0}{\Delta\omega}\right)^{\frac{3}{2}} \exp\left[-\left(\frac{\Delta\omega_0}{\Delta\omega}\right)^{\frac{3}{2}}\right]. \quad (3.8)$$

In the *quadratic* Stark effect,

$$\Delta\omega = tF^2, \quad \Delta\omega_0 = tF_0^2,$$

t being another constant. When this is substituted in (3.5) we obtain, again with the use of (3.7),

$$I(\Delta\omega) = \frac{3}{4\Delta\omega}\left(\frac{\Delta\omega_0}{\Delta\omega}\right)^{\frac{1}{2}} \exp\left[-\left(\frac{\Delta\omega_0}{\Delta\omega}\right)^{\frac{1}{2}}\right]. \quad (3.9)$$

The "normal" frequency $\Delta\omega_0$, though denoted by the same symbol, has different values in the linear and the quadratic Stark effect.

The foregoing elementary considerations, which yield the binary approximations to the intensity distribution Eqs. (3.8) and (3.9), have been superseded by the work of Holtsmark,[26] Verwey,[27] Schmaljohann,[28] and others, who included the cooperative effect of many ions. Their results are most easily derived by a method introduced by Markoff, a good presentation of which is available in a review by Chandrasekhar[5] while its application to other line-width problems is to be found in reference 2e. The treatment under review ignores the Boltzmann factor which attaches to the configuration probability of the moving ions. Effects of this simplification are considered later.

The field F, which appears in Eq. (3.5), must here be written as a vector sum when it is compounded from the fields of many ions in different places:

$$F = |\sum_i \mathbf{F}_i|. \quad (3.10)$$

The analysis leading to (3.5) must be carried out in a configuration space of $3N$ dimensions. In place of (3.5),

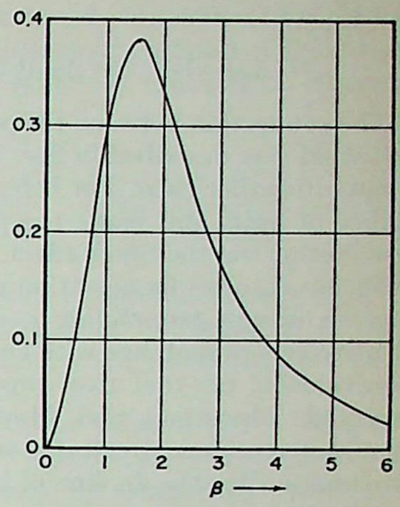


FIG. 2. Graph of $W(\beta)$ *vs* β. See Eq. (3.11b).

it yields

$$dP(F) = W(\beta)d\beta; \quad \beta = F/F_0, \quad (3.11a)$$

$$W(\beta) = \begin{cases} (4/3\pi)\beta^2(1-0.463\beta^2+0.1227\beta^4\cdots) \\ 1.496\beta^{-\frac{5}{2}}(1+5.107\beta^{-\frac{3}{2}}+14.93\beta^{-3}+\cdots). \end{cases} \quad (3.11b)$$

The two alternate expressions are useful for small and large values of β, respectively. Near $\beta=3$, both series converge slowly. According to Schmaljohann and Unsöld $W(3)=0.175$; $W(3.5)=0.122$. In Fig. 2, the function $W(\beta)$ is plotted, showing that

$$\lim_{\beta\to\infty} W = 1.496\beta^{-\frac{5}{2}} = 1.496\left(\frac{F_0}{F}\right)^{\frac{5}{2}}.$$

This is very nearly the same as the limit of Eq. (3.5),

$$\lim_{\beta\to\infty} \frac{dP(F)}{dF}\frac{dF}{d\beta} = \frac{3}{2}\left(\frac{F_0}{F}\right)^{\frac{5}{2}},$$

as asserted earlier.

These statements are not quite correct. In the binary theory F_0 was defined as

$$Ze/r_0^2 = Ze[(4\pi/3)n]^{\frac{2}{3}} = 2.60eZn^{\frac{2}{3}}.$$

The many-ion calculation involves a slightly different parameter,

$$F_0 = 2.61Zen^{\frac{2}{3}},$$

but the change is so small that its importance is quite academic, so we do not trouble to alter notation because of it.

From Eqs. (3.11) we pass to the frequency distribution in the same way as before. In the linear Stark effect $\Delta\omega/\Delta\omega_0 = F/F_0 = \beta$, and since in general

$$I(\Delta\omega)d(\Delta\omega) = W(\beta)d\beta, \quad (3.12)$$

we find

$$I(\Delta\omega) = W(\Delta\omega/\Delta\omega_0)(1/\Delta\omega_0). \quad (3.13)$$

In the *qradratic Stark effect*

$$(\Delta\omega/\Delta\omega_0)^{\frac{1}{2}} = F/F_0 = \beta,$$

[26] J. Holtsmark, Ann. Physik **58**, 577 (1919); Physik Z. **20**, 162 (1919); **25**, 73 (1924).
[27] S. Verwey, Dissertation, Amsterdam, 1936.
[28] P. Schmaljohann, Staatsexamens-Arbeit, Keil, 1936. For comments on this and reference 27, see reference 18, p. 308 ff.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

and Eq. (3.5) gives

$$I(\Delta\omega)=W[(\Delta\omega/\Delta\omega_0)^{\frac{1}{2}}]\cdot\tfrac{1}{2}(\Delta\omega_0\cdot\Delta\omega)^{-\frac{1}{2}}. \quad (3.14)$$

The connection between the preceding considerations and what was described in Sec. I as a statistical theory is not altogether clear. For here, we calculate the probability of *fields* and make the passage to an intensity distribution via the Stark effect, whereas the statistical quantum theories fix attention upon the probability of a given *energy perturbation* ϵ, a scalar quantity, and identify this probability with the line intensity at $\hbar\Delta\omega$. Equivalence of the two procedures is intuitively expected. Margenau and Meyerott,[11a] who made a quantum mechanical calculation using the binary approximation for the L_α line of hydrogen, indicate how the equivalence comes about mathematically. Among other things ϵ is not the sum, $\sum_i \epsilon_i$ for individual perturbers but is composed very much like $\mathbf{F}$ from the vectors $\mathbf{F}_i$; hence the Holtsmark theory does not depart from the statistical pattern.

This treatment includes the effect of the nonuniformity of the Stark field produced by ions. In the ordinary theory of the Stark effect, $\mathbf{F}$ is treated as a constant, whereas the ions produce a field which depends on r, the dependence being greater the smaller r. In general the nonuniformity effect is small, increasing in magnitude with ion density, $W(\beta)$ as given by (3.11b) no longer has the asymptotic form $1.5\,\beta^{-\frac{5}{2}}$, but $1.5\,(\beta^{-\frac{5}{2}}+2\beta^{-2}a'/r_0)$, a' being the radius of the lower atomic state (first Bohr radius for L_α). The nonuniformity also causes shifts and widths in Stark components which do not show a first-order effect, the "forbidden" half-widths being of the order a'/r_0 times the regular Holtsmark widths.

B. Effect of Perturber Interactions

In part (A) of this section the probability of a given ion configuration was taken to be proportional to the volume of configuration space assignable to it, the Boltzmann factor $\exp[-V(\mathbf{r}_1,\mathbf{r}_2\cdots,\mathbf{r}_N)/kT]$ being neglected. This factor confers smaller weight upon configurations in which several ions are close together and therefore reduces the probability that many ions shall be situated near the radiating atom. Hence the Holtsmark theory overestimates the likelihood of strong perturbations and therefore the intensity in the wings of a line.

An accurate calculation with inclusion of the Boltzmann factor has not been made. Broyles[29] considers the problem most carefully, but his results are relevant primarily for the case where the radiator, too, is ionized and repels the perturbing ions by a Coulomb field of its own. His results are examined in the next section. Here we summarize a contribution by Ecker,[30] which deals directly with the problem. He relies upon the Debye-Hückel screening mechanism to provide a semiquantitative solution, which introduces the electrons as well as the ions into the picture.

In the presence of a cloud of ions and electrons the Coulomb field produced by an ion is given by

$$\mathbf{F}=(Ze\mathbf{r}/r^2)[(1/r)+(1/D)]e^{-r/D},$$

as may be seen by differentiating the expression for V, Eq. (5.4). According to (5.5),

$$D=[kT/4\pi ne^2(1+Z^2)]^{\frac{1}{2}}.$$

Analysis of the line-broadening problem even with this simplified potential remains formidable; Ecker chooses as an approximation to $\mathbf{F}$ the form§

$$\mathbf{F}=\begin{cases} Ze^2\mathbf{r}/r^3 & \text{if } r<D \\ 0 & \text{if } r\geq D. \end{cases}$$

Machine computation by the method of part (A) leads to a dependence of the line shape on the parameter

$$\delta=(4\pi D^3/3)n,$$

whose physical meaning is the number of ions within the Debye radius D. Clearly, if $D\to\infty$, the result must agree with Holtsmark's. Figure 3 shows the results of Ecker's calculations for three values of δ. The expected effect, reduction of the intensity in the wing is clearly in evidence.

In principle, a further correction must be made in the Holtsmark distribution. The Debye-Hückel cutoff modifies the effective field at large r. For small r, the Coulomb field breaks down because of *atomic* screening, the potential going to a finite value as $r\to 0$. The theory of this effect is easily developed[31] and predicts a flattening of the Holtsmark curve far from the center of the line. In the region affected, the frequency is pro-

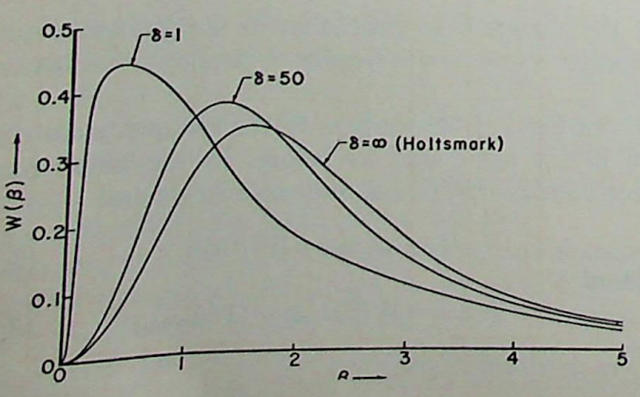


FIG. 3. Graph of $W(\beta)$ for different values of δ.

[29] A. A. Broyles, Phys. Rev. **100**, 1181 (1955).

[30] (a) G. Ecker, Z. Physik **148**, 593 (1957); (b) **149**, 254 (1957); (c) Z. Naturforsch. **12**, 346, 517 (1957).

§ A calculation of the shielding correction which avoids this cutoff was made by H. Hoffman and O. Theimer, Astrophys. J. **127**, 477 (1958). Ecker (private communication) has performed a careful machine calculation based on the correct potential and obtains results in substantial agreement with his former approximation and differing from those of Hoffman and Theimer.

[31] H. Margenau, Progress Report AF-18 (603)–15, January 1958.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

portional approximately to β^{-1}, not to $\beta^{-\frac{1}{2}}$. But this region lies so far outside the measured range of line widths (hundreds of angstroms for the Balmer lines, as a simple estimate, based on the assumption that the breakdown radius is of atomic dimensions, shows) that this correction is at this time academic.

C. Treatment for Radiating Ions

The analysis leading to Eq. (3.5) is further defective in assuming no potential energy between the moving ions, and none between the radiating atom and the ions. The former is always present but has ordinarily no very profound effect on the frequency distribution. The second potential energy is present when the radiator itself is ionized; in that case its effect is more important. Attention has been called to these shortcomings by Mayer[32] and Broyles,[29] whose interest is primarily in highly ionized exploding plasmas where the ions themselves emit the lines. A completely adequate theory of these effects is not at hand. We develop here first the binary approximation to the many-ion analysis and compare the results with Eq. (3.4). (Our development completes somewhat the treatment of the aforenamed authors who omitted the condition that there be no ion within r and obtained an unnormalized distribution.)

The perturbing ion of charge Z_2e will repel the radiating ion of charge Z_1e with the Coulomb energy $Z_1Z_2e^2/r$. The former p_- will then be $1-4\pi ne^{-a/r}r^2dr$, provided

$$a=(Z_1Z_2e^2)/(kT). \tag{3.15}$$

In place of (3.2) we obtain

$$dP(a,r)=4\pi n\,\exp[(-a/r)-4\pi nA(a,r)]r^2dr, \tag{3.16}$$

where

$$A(a,r)=\int_0^r e^{-a/r}r^2dr. \tag{3.17}$$

Hence, the ratio of (3.16) to the Holtsmark distribution (3.4) is

$$S=dP(a,r)/dP(0,r)$$
$$=\exp\{-(a/r)+4\pi n[A(0,r)-A(a,r)]\}. \tag{3.18}$$

The function $A(a,r)$ is easy to calculate; it may be written

$$A(a,\xi a)=\tfrac{1}{3}a^3\mathfrak{F}(\xi);$$
$$\mathfrak{F}(\xi)\equiv(\xi^3-\tfrac{1}{2}\xi^2+\xi)e^{-1/\xi}+\tfrac{1}{2}Ei(-1/\xi). \tag{3.18a}$$

the function $\mathfrak{F}$ is plotted in Fig. 4. Clearly,

$$A(0,r)=r^3/3.$$

For the binary approximation to be valid, the ratio $(r_0/r)^2=\beta$ must be greater than about 7. On the other hand, $r_0=0.62n^{-\frac{1}{3}}$. The last results, then, begin to be useful at values of r in the neighborhood of $1/4n^{-\frac{1}{3}}$ and remain so for smaller radii. At this critical distance, Eq. (3.18) becomes

$$S=\exp\left\{-4an^{\frac{1}{3}}+\frac{\pi}{48}-\frac{4\pi}{3}na^3\mathfrak{F}\left(\frac{1}{4an^{\frac{1}{3}}}\right)\right\}. \tag{3.19}$$

The terms in the exponent become important when $an^{\frac{1}{3}}\approx 1$. This is the condition for equality of mean kinetic energy and Coulomb energy at the mean distance of separation of the ions.

If $4an^{\frac{1}{3}}=0.1$, $S=e^{-0.09}$, the value of $\mathfrak{F}(10)$ being 845. For $4an^{\frac{1}{3}}=1$, $S=e^{-0.952}$ (since $\mathfrak{F}(1)=0.258$), and for larger values of $4an^{\frac{1}{3}}$ only the term $\exp(-4an^{\frac{1}{3}})$ of S remains, the others being negligible.

In the light of these numerical results we consider two cases, first the perturbation of singly ionized helium atoms by protons in a star or a discharge plasma. Here $Z_1=Z_2=1$. The condition that the binary approximation be significant, namely $r<0.62n^{-\frac{1}{3}}$, is easily compatible with the requirement that r be greater than a helium ion, $r>5\times10^{-9}$ cm, since these inequalities merely imply $n<10^{24}$ cm^{-3}. But to have S appreciably smaller than 1, $an^{\frac{1}{3}}$ must be near 1. This means that $T\approx(e^2/k)n^{\frac{1}{3}}=1.7\times10^{-3}n^{\frac{1}{3}}$ °K, which can occur in a shock front or in an arc.

The situation is quite different for atomic explosions. At amosphere of iron ions, each with $Z=23$, and with $T\approx10^7$ °K implies $a=8.4\times10^{-8}$ cm. At normal density

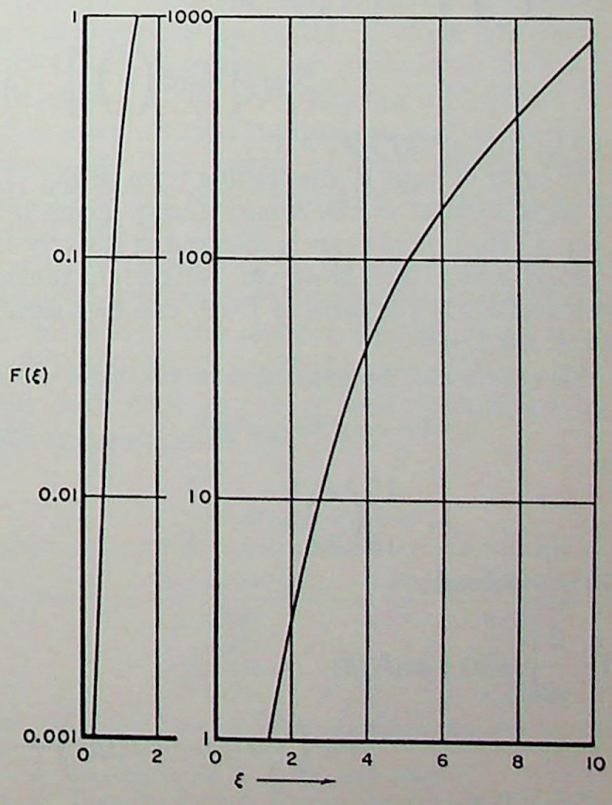


FIG. 4. Graph of $F(\xi)$ *vs* ξ. See Eq. (3.18a).

[32] H. Mayer, Los Alamos Scientific Laboratory Report LA-647, 1947.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

$n \approx 10^{23}$ cm^{-3} and the condition upon r is $r < 10^{-8}$ cm. Because of the small size of the iron core this leaves a considerable range in which the binary approximation is correct. Indeed at $r = 1/4n^{-\frac{1}{3}}$, for which (3.19) was computed $S \approx 10^{-15}$, indicating that the Holtsmark formula is enormously in error.

We now treat the many-ion problem[33] using the following model: the free electrons form a uniformly smeared out negative charge, and interactions are considered only between the radiating ion and the plasma ions; interaction between plasma ions themselves is ignored. This method avoids some uncertainties inherent in the approximations using the Bohm-Pines approach (see the following), for the approximations are clear in their physical meaning and the analysis is otherwise exact.

With this model, the probability $P(\mathbf{F})$ for an electric field F is given by

$$P(\mathbf{F}) = \frac{\int \cdots \int \exp(-a \sum_{i \to N} r_i^{-1}) \delta_i(\mathbf{F} - \sum_{i \to N} Q_2 \mathbf{r}_i / r_i^3) d\mathbf{r}_1 \cdots d\mathbf{r}_N}{\int \cdots \int \exp(-a \sum_i r_i^{-1}) d\mathbf{r}_1 \cdots d\mathbf{r}_N}, \tag{3.20}$$

where a is defined by Eq. (3.15). Equation (3.20) can be written in a more convenient form, best obtained through use of Chandrasekhar's method.[5] As before, we denote by $W(\beta)$ the probability of finding a field of strength $\beta = F/F_0$ [see Eq. (3.11a)]

$$F_0 = 2.61 Z_2 e n^{\frac{2}{3}}.$$

F_0 is the quantity introduced previously and represents essentially the field produced on one ion by another ion at the average spacing between ions, and n is the number density of ions. Then

$$W(\beta) = \frac{2}{\pi\beta} \int_0^\infty \exp\left[-\left(\frac{x}{\beta}\right)^{\frac{3}{2}} \eta\right] x \sin x dx, \tag{3.21}$$

$$\eta = \frac{15}{8}\left(\frac{2}{\pi}\right)^{\frac{1}{2}} \int_0^\infty dz z^{-7/2}(z - \sin z) \times \exp\left[-\alpha z^{\frac{1}{2}} \left(\frac{\beta}{x}\right)^{\frac{1}{2}}\right], \tag{3.22}$$

and $\alpha = [Z_1 e (Z_2 e)^{\frac{1}{2}} (kT)] F_0^{-\frac{1}{2}}$.

In the limit of large β, the leading term in Eq. (3.21) is the same as that of the binary theory given before [see Eq. (3.16)]. $W(\beta)$ can be computed directly from Eqs. (3.21) and (3.22). However, for certain ranges of α ($\alpha \ll 1$ and $\alpha \gg 1$) the form of $W(\beta)$ can be simplified.

Case 1: $\alpha \ll 1$, $\alpha\beta^{\frac{1}{2}} \ll 1$.

In this case η can be expanded in the ratio $y = x/\beta$:

$$\eta = 1 - c_1 \alpha y^{-\frac{1}{2}} + c_2 \alpha^2 y^{-1} \cdots, \tag{3.23}$$

where

$$c_1 = \frac{15}{16}\left(\frac{\pi}{2}\right)^{\frac{1}{2}}, \quad c_2 = \frac{5}{4}.$$

$W(\beta)$ then becomes

$$W(\beta) = \frac{2}{\pi\beta}\left[I_0(\beta) + \alpha c_1 I_1(\beta) + \alpha^2\left(\frac{c_1^2}{2} I_2(\beta) - c_2 I_{\frac{3}{2}}(\beta)\right) + \cdots\right],$$

where

$$I_p(\beta) = \int_0^\infty \exp\left[-\left(\frac{x}{\beta}\right)^{\frac{3}{2}}\right] x \left(\frac{x}{\beta}\right)^p \sin x dx. \tag{3.24}$$

It is seen that $(2/\pi\beta) I_0(\beta)$ is the Holtsmark distribution, to which Eq. (3.24) reduces in the limit, $T \to \infty$ ($\beta \to 0$).

The integrals in Eq. (3.24) can be expanded in a series and evaluated for $\beta \leqslant 3$ and $\beta \geqslant 6$. The range from 3 to 6 is difficult because of the slow convergence of the series. The coefficients of the first two terms in Eq. (3.24) are listed in Table III.

Case 2: $\alpha \gg 1$, $\beta \ll \alpha$.

For this case, η is expanded in an asymptotic series in y:

$$\eta = \frac{15}{4}\left(\frac{2}{\pi}\right)^{\frac{1}{2}}\left(\frac{1}{3!} y^{\frac{3}{2}} \alpha^{-1} - \frac{4!}{5!} y^{\frac{5}{2}} \alpha^{-5} + \frac{8!}{7!} y^{9/2} \alpha^{-9} - \cdots\right). \tag{3.25}$$

If only the first term is retained, one finds

$$W(\beta) = \frac{\alpha^{\frac{3}{2}} \beta^2 \exp(-\beta^2 \alpha (2\pi)^{\frac{1}{2}}/5)}{2\pi^{\frac{1}{2}}[(5/8)(2/\pi)^{\frac{1}{2}}]^{\frac{3}{2}}} = \frac{\alpha^{\frac{3}{2}} \beta^2 \exp[-0.5\alpha\beta^2]}{1.25}. \tag{3.26}$$

Equation (3.26) is identical with a formula proposed by Mayer[32] on the basis of a different, simpler physical

TABLE III. Coefficients of the first two terms of Eq. (3.24). Values of $(2/\pi\beta) I_0(\beta)$ from reference 5.

β	$(2/\pi\beta) I_0(\beta)$	$(2/\pi\beta) c_1 I_1(\beta)$
0.1	0.004225	0.00745
0.6	0.129598	0.21264
1.0	0.271322	0.3860
2.0	0.33918	0.1791
3.0	0.176	−0.08707
6.0	0.02417	−0.0507
8.0	0.01038	−0.0273
10.0	0.00556	−0.0168

[33] M. B. Lewis and H. Margenau, Phys. Rev. 109, 842 (1958).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

model. Why the two models lead to the same result is not clear.

Broyles[29] has proposed another way of computing the probability $P(F)$ of finding an electric field of magnitude F at a radiating *ion* in a plasma. He uses the method of Pines and Bohm[34] to separate the potential energy and the electric field into a short and a long-range component.

The work is based on a model in which the free electrons are represented as a uniformly smeared out negative charge. The following units are used: unit length is the radius of a sphere whose volume is the volume per ion r_0; unit field strength is F_0, and unit energy is that of two ions separated by unit distance, and the temperature is expressed in the above energy units.

The probability density for occurrence of a field β is [see Eq. (3.11a) for definition of β] in analogy to Eq. (3.20),

$$P(\beta)=\frac{\int\cdots\int \exp(-V/\theta)\delta(\beta-\sum_i \mathbf{r}_i/r_0^3)d\mathbf{r}_1\cdots d\mathbf{r}_N}{\int\cdots\int \exp(-V/\theta)d\mathbf{r}_1\cdots d\mathbf{r}_N}, \tag{3.27}$$

provided V is the potential energy, and δ the Dirac "function." On using the Dirichlet representation for δ one obtains

$$P(\beta)=(2\pi)^{-3}\int T(\mathbf{l})\exp(i\beta\cdot\mathbf{l})d\mathbf{l}, \tag{3.28}$$

$$T(\mathbf{l})\equiv\frac{\int\cdots\int \exp[-(V/\theta)+i\mathbf{l}\cdot\sum_j \mathbf{r}_j/r_j^3]d\mathbf{r}_1\cdots d\mathbf{r}_N}{\int\cdots\int \exp(-V/\theta)d\mathbf{r}_1\cdots d\mathbf{r}_N}.$$

Since $P(\beta)$ depends only on the magnitude of β we have

$$W(\beta)=4\pi\beta^2P(\beta). \tag{3.29}$$

We now employ the Pines-Bohm method to split the potential energy and the electric field into short and long-range components. The *ion* charge density ρ_i is written as

$$\rho_i=\sum_j \delta(\mathbf{r}-\mathbf{r}_j). \tag{3.30}$$

But the total charge density ρ can be expanded into a Fourier series (unit volume per particle) to give

$$\rho=\sum_k{}' \rho(\mathbf{k})\exp(i\mathbf{k}\cdot\mathbf{r}), \tag{3.31}$$

with

$$\rho(\mathbf{k})=\sum_j \exp(-i\mathbf{k}\cdot\mathbf{r}_j),$$

where the prime indicates the omission of $k=0$ from the sum in order that the net charge be zero. The potential and the electric field when expressed as functions of k can each be written as a sum of two terms, one for high and one for low values of k. The high k terms are functions of the individual $\mathbf{r}_i$ and are referred to as the particle or short-range part, the low k terms are expressible as functions of the $\rho(\mathbf{k})$ and are called the wave or long-range part of the respective quantity. With this separation, $T(\mathbf{l})$ can likewise be approximated in consistent fashion; it becomes a product of T_w (wave parts) and T_p (particle part). The wave part can easily be evaluated; the particle part is approximated by Broyles in two ways:

(1) Short-range central interactions (SRCI). In this approximation only the interactions between the radiating ion and the perturbing ions are retained, the interactions among the perturbing ions being neglected. This approximation is similar to the one in the previous section, except that it was there applied to the full Coulomb interactions of the ions whereas it is here restricted to the short range components which arise in the Bohm-Pines method.

(2) Short-range nearest neighbor (SRNN) interactions. Interactions are considered only between the radiating ion and its nearest neighbor.

Broyles[29] has employed his method for iron ions at normal densities, with a charge of 23 electron charges at a temperature of 1 kev, a case already considered earlier in this section. This corresponds to $\theta=0.186$. Under these conditions, α as defined in the previous part of this section [Eq. (3.22)] has a value of 5.5. It is therefore possible to use Eq. (3.26) for $\beta\ll 5.5$. The error attending the use of Eq. (3.26) is less than 10% for $\beta=1$ and decreases for decreasing β. Table IV compares Eq. (3.26) with Broyles' SRNN for $0.3\leqslant\beta\leqslant 1.5$ A more detailed comparison is given in Fig. 2 of reference 29. The curve marked "simple harmonic oscillator" is the result of Mayer's calculation and is identical with Eq. (3.26).

A separate publication by Broyles[35] deals extensively with the approximations involved in using the Pines-Bohm method for this problem and devises an improved procedure[36] which is applicable for $\theta\geqslant 0.6$. It is concluded that $P(\beta)$ is rather well determined by the SRNN approximation for $\theta\geqslant 0.6$.

TABLE IV.

β	Eq. (3.26) $\alpha=5.5$	Broyles' SRNN; $\theta=0.186$
0.3	0.69	0.63
0.5	1.25	1.13
1	0.64	0.66
1.5	0.04	0.147

[34] D. Pines and D. Bohm, Phys. Rev. **85**, 338 (1952).

[35] A. A. Broyles, Phys. Rev. **105**, 347 (1957).

[36] A. A. Broyles, Atomic Energy Commission Report R.M.—1682.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

IV. REFINED IMPACT THEORY. FUSION OF IMPACT AND STATISTICAL THEORIES

A. Impact Theory

A consistent and ingenious development of the impact method was carried out by E. Lindholm[8c,15b] who was preceded in several important respects by W. Lenz[8a] and G. Burkhardt.[37] In deriving Lindholm's results we use recent stochastic methods.

Let us return to Eq. (1.13), which represents the true line form, provided only the adiabatic hypothesis holds [see formula (1.11)]. The immediate problem is to calculate the correlation function C. To strip it of irrelevancies, we continue to suppose that the radiator has only two states, and that only one of these is occupied: $S_1=1$, $S_2=0$. We also omit subscripts. Then,

$$C(\tau)=\mu^2\int_{-\infty}^{\infty}e^{-i[\varphi(t+\tau)-\varphi(t)]}dt. \tag{4.1}$$

The phase φ is defined by (1.12). In that expression, E_l and E_m contain constant parts, E_l^0 and E_m^0, plus the perturbations. The constant parts contribute to φ_{lm} the amount $[(E_l^0-E_m^0)/\hbar]t\equiv\omega' t$, ω' being the normal frequency of the line. In this section, *we redefine φ (without changing notation) to mean only the variable part which is contributed by the perturbations*. As a result, we change Eq. (1.13a) to read

$$I(\omega)\propto\int d\tau e^{i(\omega-\omega')\tau}C(\tau). \tag{4.2}$$

Positive φ corresponds to a positive difference ΔE (upper state)$-\Delta E$ (lower state). There is some confusion on this point in the cited literature.

The quantity $\varphi(t+\tau)-\varphi(t)$ represents the perturbational phase change which has occurred in the interval $(t, t+\tau)$. The integrand in Eq. (4.1) is a time series in which this quantity changes from instant to instant t, and we are integrating over t. As is customary (ergodic hypothesis) the one time series is replaced by an ensemble of time series and the *time integral* is identified with an *ensemble average* (except for a constant factor). Thus

$$C(\tau)\propto\langle e^{-i\varphi(\tau)}\rangle_{Av}\equiv\int_{-\infty}^{\infty}e^{-i\varphi}P_\tau(\varphi)d\varphi, \tag{4.3}$$

where $P_\tau(\varphi)$ is the probability of a phase change φ in the interval τ. It is convenient to make the assumption characteristic of the impact approach, namely that the collisions are sudden and occur singly. The value of φ is then the sum of a whole number of individual phase changes φ_i, each associated with one collision. By the rules of probability, $P_\tau(\varphi)$ is related to the probability $P_\tau(k)$ that k collisions have taken place in τ, by

$$P_\tau(\varphi)=\sum_{k=0}^{\infty}P_\tau(k)\sum_{1,2\cdots k}\left[\prod_{j=1}^{k}P(\varphi_j)\right]\delta\left(\varphi-\sum_{j=1}^{k}\varphi_j\right) \tag{4.4}$$

and

$$P_\tau(k)=(\tau\nu_c)^k e^{-\nu_c\tau}/k!, \tag{4.5}$$

which is Poisson's law. As before, ν_c is the collision frequency $P(\varphi_j)$ the probability for the occurrence of a phase change of magnitude φ_j, and δ is the well-known Dirac "function" which may be replaced by its Dirichlet representation

$$\delta\left(\varphi-\sum_{j=1}^{k}\varphi_j\right)=\frac{1}{2\pi}\int_{-\infty}^{\infty}dy\exp\left[iy\left(\varphi-\sum_{1}^{k}\varphi_j\right)\right]. \tag{4.6}$$

The symbol $\sum_{1,2\ldots k}$ is meant as a summation over all $\varphi_1, \varphi_2, \cdots, \varphi_k$ in every factor that follows it. With the aid of (4.5) and (4.6), one may write Eq. (4.4) in the form

$$P_\tau(\varphi)=\frac{e^{-\nu_c\tau}}{2\pi}\int_{-\infty}^{\infty}dy e^{iy\varphi}\sum_{k=0}^{\infty}\frac{\nu_c\tau}{k!}[\sum_i P(\varphi_j)e^{-iy\varphi_j}]^k,$$

and this may be expressed as

$$P_\tau(\varphi)=\frac{e^{-\nu_c\tau}}{2\pi}\int_{-\infty}^{\infty}dy e^{iy\varphi+\nu_c\tau\alpha(y)}, \tag{4.7}$$

provided

$$\alpha(y)=\sum_i P(\varphi_j)e^{-iy\varphi_j}. \tag{4.8}$$

On substituting (4.7) in (4.3) we have

$$C(\tau)=\frac{e^{-\nu_c\tau}}{2\pi}\int_{-\infty}^{\infty}e^{-i\varphi}d\varphi\int_{-\infty}^{\infty}dy e^{iy\varphi+\nu_c\tau\alpha(y)}$$

$$=e^{-\nu_c\tau}\int dy e^{\nu_c\tau\alpha(y)}\left\{\frac{1}{2\pi}\int_{-\infty}^{\infty}e^{i(y-1)\varphi}d\varphi\right\}.$$

The factor { } is $\delta(y,1)$. Hence,

$$C(\tau)=e^{-\nu_c\tau[1-\alpha(1)]}. \tag{4.9}$$

This result is sufficient for many purposes and underlies most of the work on impact broadening. A generalization which allows the collision probability to depend on velocity, and which also gets rid of the usual assumption that the collisions are sudden and occur singly, is contained in Kolb's[22,38] report and is given below, following the lines of the preceding account. The result [Eq. (4.16)] is used in Sec. V.

Let us divide the time interval $(-\infty, \infty)$ into elements Δt_i. The phase change produced by *one* perturber in the time interval τ depends on the time of closest approach, the velocity v, and in general on some impact parameter l. Let us call this phase change $\varphi_{i,v,l}(\tau)$,

[37] G. Burkhardt, Z. Physik 115, 592 (1940).

[38] A. C. Kolb, AFOSR-TN-57-8, Astia Document No. AD115-400.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

where i refers to the time interval $\Delta\tau_i$ which contains the time of closest approach. If the perturbations are *scalarly additive*, the total phase change is

$$\varphi(\tau)=\sum_{iv}\varphi_{iv}(\tau),$$

where $\varphi_{iv}(\tau)$ is the phase change produced in the time interval τ by all perturbers described by (i,v). Clearly,

$$\varphi_{iv}(\tau)=\sum_{j=1}^{\mathfrak{N}(iv)}\varphi_{i,v,l_j}(\tau),$$

if $\mathfrak{N}(iv)$ is the number of perturbers characterized by (i,v) in some effective volume. We first compute $P(\varphi_{iv})$, the probability of a phase change φ_{iv} due to all perturbers in the class (i,v).

Assume that the effective volume of our system is a sphere of radius R and write $n(v)$ for the number per cubic centimeter of perturbers with velocity v (assumed at first to take on discrete values). We designate by $W(l)$ and $W[\mathfrak{N}(i,v)]$ the probability distributions for l and $\mathfrak{N}(i,v)$. The latter is given again by Poisson's distribution,

$$W[\mathfrak{N}(i,v)]=\frac{1}{\mathfrak{N}(i,v)!}(\nu_v\Delta t_i)^{\mathfrak{N}(iv)}e^{-\nu_v\Delta t_i}, \quad (4.10)$$

where ν_v is the collision frequency for particles with velocity v and is given by

$$\nu_v=\pi R^2 v n(v). \quad (4.11)$$

We then have

$$P(\varphi_{i,v})=\sum_{\mathfrak{N}(iv)=0}^{\infty}\int_{l_1}\cdots\int_{l_{\mathfrak{N}(i,v)}}W[\mathfrak{N}(i,v)]\times\left\{\prod_{j=0}^{\mathfrak{N}(i,v)}W(l_j)dl_j\right\}\delta\left(\varphi_{iv}-\sum_{j=1}^{\mathfrak{N}(iv)}\varphi_{i,v,l_j}\right). \quad (4.12)$$

Again we replace the δ function by its Dirichlet representation

$$\delta(x-x')=\frac{1}{2\pi}\int_{-\infty}^{\infty}dy e^{iy(x-x')}, \quad (4.13)$$

substitute (4.10) into (4.12) and obtain

$$P(\varphi_{iv})=\frac{1}{2\pi}\int_{-\infty}^{\infty}dy\exp\{iy\varphi_{iv}+\nu_v\Delta t_i[\alpha_{i,v}(y)-1]\}, \quad (4.14)$$

provided

$$\alpha_{i,v}(y)=\int W(l)e^{-iy\varphi_{iv}l}dl.$$

Now $P(\varphi)$ can be written

$$\int\cdots\int\left(\prod_{iv}P(\varphi_{iv})d\varphi_{iv}\right)\delta\left(\varphi-\sum_{iv}\varphi_{iv}\right).$$

On using Eqs. (4.13) and (4.14) this becomes, after rearrangement,

$$P(\varphi)=\frac{1}{2\pi}\int_{-\infty}^{\infty}e^{is\varphi}ds\left\{\prod_{iv}\int_{-\infty}^{\infty}dy e^{\nu_v\Delta t_i[\alpha_{i,v}(y)-1]}\cdot\frac{1}{2\pi}\int d\varphi_{iv}e^{i\varphi_{i,v}(y-s)}\right\}.$$

The last integral is $\delta(y-s)$, so that

$$P(\varphi)=\frac{1}{2\pi}\int_{-\infty}^{\infty}e^{is\varphi}ds\prod_{iv}e^{\nu_v\Delta t_i[\alpha_{iv}(s)-1]}=\frac{1}{2\pi}\int_{-\infty}^{\infty}\exp\{is\varphi+\sum_{i,v}\nu_v\Delta t_i[\alpha_{i,v}(s)-1]\}ds.$$

$C(\tau)$ then becomes proportional to

$$\int_{-\infty}^{\infty}ds\exp\{\sum_{iv}\nu_v\Delta t_i[\alpha_{i,v}(s)-1]\}\cdot\frac{1}{2\pi}\int_{-\infty}^{\infty}d\varphi e^{i(s-1)\varphi}$$

and, therefore,

$$C(\tau)\propto\exp\{\sum_{i,v}\nu_v\Delta t_i[\alpha_{i,v}(1)-1]\}.$$

This finally can be written with the use of Eq. (4.4) and replacement of the sum by integrals, as

$$C(\tau)\propto\exp\left\{\int dv\int_{-\infty}^{\infty}dl\pi R^2 v n(v)[\alpha_{l,v}(1)-1]\right\}. \quad (4.15)$$

We now apply this general formula to the special case of impacts. In this approximation, all collisions are assumed to be short enough so that those having their time of closest approach in τ are also completed in τ, and one may disregard the others. This means, in terms of $\varphi_{t,v,l}$, that

$$\varphi_{t,v,l}\begin{cases}=0 \text{ if } t \text{ is not in } \tau\\ =\varphi_{v,l} \text{ if } t \text{ is in } \tau.\end{cases}$$

In the second instance $\varphi_{t,v,l}$ is independent of time, and therefore $\varphi_{v,l}$ represents the total phase change produced by *one* particle with time of closest approach in τ and with impact parameter l. Under these conditions

$$\alpha_{t,v}(1)=1 \text{ if } t \text{ is not in } \tau$$

$$\alpha_{t,v}(1)=\alpha_v(1)=\int dl W(l)e^{-i\varphi_{vl}} \text{ if } t \text{ is in } \tau.$$

Equation (4.15) then leads to

$$C(\tau)\propto\exp\left\{\pi R^2\int dv v n(v)[\alpha_v(1)-1]\tau\right\}. \quad (4.16)$$

This is the general form of $C(\tau)$ for an impact theory

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

based on scalarly additive perturbations. We return to this equation later (Sec. V). Equation (4.16) reduces to (4.9) under the proper conditions. To show this we first assume that all perturbers have the same velocity v'. This means we replace $n(v)$ by $n\delta(v-v')$, where n is the number of particles per cubic centimeter. In view of Eq. (4.11) and because ν_v is now identical with our former ν_c, we then recover Eq. (4.9)

$$C(\tau) \propto e^{\nu_c[\alpha(1)-1]\tau}.$$

This is the correlation function for perturbers all of which have the same velocity. This derivation shows that its validity rests on the premise of scalarly additive perturbations and does not require the restriction that the impacts fail to overlap in time; this feature was not evident from the simpler proof which led to Eq. (4.9).

We return to calculation of the line intensity, employing the correlation function given by Eq. (4.9). We split $\alpha(1)$ into its real and its imaginary part, $\alpha=\alpha_1+i\alpha_2$ and put

$$\nu_c(1-\alpha_1)\equiv u_1, \quad -\nu_c\alpha_2\equiv u_2 \tag{4.17}$$

so that

$$C(\tau)=e^{-(u_1+iu_2)\tau}. \tag{4.18}$$

The calculation of I in accordance with Eq. (4.2) is then easy. One further detail has to be remembered. It follows from (4.1) that $C(-\tau)=C^*(\tau)$, a relation necessary for the reality of $I(\omega)$ but obscured by the explicit form (4.18) which is valid for $\tau>0$. Thus

$$C(\tau)=C^*(-\tau)=e^{(u_1-iu_2)\tau} \quad \text{for } \tau<0$$

and by (4.2)

$$I(\omega)\propto\left[\int_{-\infty}^{0} e^{u_1\tau+i(\omega-\omega'-u_2)\tau}d\tau+\int_{0}^{\infty} e^{-u_1\tau+i(\omega-\omega'-u_2)\tau}d\tau\right]$$
$$=2u_1[u_1^2+(\omega-\omega'-u_2)^2]^{-1}.$$

To make $\int_0^\infty I(\omega)d\omega=1$ we must write

$$I(\omega)=u_1/\pi[u_1^2+(\omega-\omega'-u_2)^2]^{-1}. \tag{4.19}$$

This represents a dispersion curve with a shift

$$u_2=\nu_c\sum_j P(\varphi_j)\sin\varphi_j \tag{4.20}$$

to the blue of the normal peak at ω', and a half-width

$$\omega_{\frac{1}{2}}=2u_1=2\nu_c[1-\sum_j P(\varphi_j)\cos\varphi_j]. \tag{4.21}$$

Lindholm now falls back upon Eq. (1.5) for a computation of the φ_j, which he identifies with the $\Delta\varphi(\rho,v)$ of that expression. This proficient and simple step is warranted under two conditions: (*a*) The impacts are sudden and do not coincide; (*b*) the perturber has a classical path, so that $\epsilon[(\rho^2+v^2t^2)^{\frac{1}{2}}]$ is meaningful. The first of these had already been discussed; the second is an additional restriction to be examined in the light of Sec. II.

Equation (4.21) distinguishes between phase changes by the arbitrary index j; Eq. (1.5) classifies them by the impact parameter ρ which is the closest distance of approach. The connection is this: If φ_j is taken to be $\varphi(\rho)$ then $P[\varphi(\rho)]=2\pi\rho d\rho/q$, where q is the total collision cross section. On the other hand, $\nu_c=nqv$. Therefore (4.20) and (4.21) become

$$u_2=nv\int 2\pi\rho d\rho\,\sin\Delta\varphi(\rho,v), \tag{4.22}$$

$$\omega_{\frac{1}{2}}=2nv\left[q-\int 2\pi\rho d\rho\,\cos\Delta\varphi(\rho,v)\right]$$
$$=2nv\int(1-\cos\Delta\varphi)2\pi\rho d\rho=4\pi nv\int\sin^2\Delta\varphi\rho d\rho. \tag{4.23}$$

The integrations over ρ are to be carried from 0 to $(q/\pi)^{\frac{1}{2}}$. Because of the rapid decline of $\Delta\varphi$ with ρ it is often permissible to replace the upper limit by ∞. In the case of long-range Coulomb interactions there is the added difficulty of giving a meaning to q.

The evaluation of $\Delta\varphi$ proceeds in accordance with (1.4); it is profitable only for simple force laws of the form

$$\epsilon(r)=\hbar\Omega_\sigma/r^\sigma. \tag{4.24}$$

Here we have replaced the former C_σ by $\hbar\Omega_\sigma$ to save writing. For our purposes the form (4.24) is satisfactory, for we shall be dealing mainly with linear Stark effects (where $\epsilon=\hbar sF=\hbar se/r^2$, whence $\sigma=2$) and quadratic Stark effects (where $\epsilon=\hbar tF^2=\hbar te^2/r^4$, whence $\sigma=4$). Here ϵ is the difference in the Stark displacements of the two states; it refers to the line, not to a given state.

With (4.24) Eq. (1.5) yields

$$\Delta\varphi=\Omega_\sigma\int_{-\infty}^{\infty}dt(\rho^2+v^2t^2)^{-\sigma/2}=\pi^{\frac{1}{2}}\frac{\Gamma[(\sigma-1)/2]}{\Gamma(\sigma/2)}\frac{\Omega_\sigma}{v\rho^{\sigma-1}}.$$

In particular, for $\sigma=2$

$$\Delta\varphi=(\pi\Omega_2)/(v\rho) \tag{4.25}$$

and for $\sigma=4$

$$\Delta\varphi=(\pi\Omega_4)/(2v\rho^3). \tag{4.26}$$

We now compute (4.22) and (4.23) for these two cases.

The shift for the linear Stark effect is not observable as such because the splitting pattern as a whole is symmetric. Lindholm and Unsöld do not compute it for that reason. Nevertheless, it is meaningful as a measure of the shift of each Stark component, and it does manifest itself in each wing of such lines as H_β and H_δ, which have zero intensity at the center. Since q is not well defined, the upper limit in the integrations of (4.22) and (4.23) will be taken as r_0 (see Sec. III: $(4\pi/3)r_0^3n=1$), this being the order of distance beyond which the

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

Coulomb force is effectively screened.‖ When $\sigma=4$ the convergence is fast and the integrations may be carried to ∞. For $\sigma=2$ both integrations are straightforward (see Lindholm); the forms below involve an expansion of the integral sine and cosine. For $\sigma=4$ a graphical integration is necessary to determine the numerical factors. Thus one finds:

Linear Stark effect, $\sigma=2$

$$\text{Shift}=u_2=2\pi^3\Omega_2^2\frac{n}{v}\left(\frac{vr_0}{\pi\Omega_2}-\frac{\pi}{4}+\frac{\pi\Omega_2}{2vr_0}-\cdots\right), \quad (4.27)$$

$$\text{Half-width}=\omega_{\frac{1}{2}}=2\pi^3\Omega_2^2\frac{n}{v}\left(0.923-\ln\frac{\pi\Omega_2}{vr_0}-\frac{\pi^2\Omega_2^2}{24v^2r_0^2}+\cdots\right). \quad (4.28)$$

Quadratic Stark effect, $\sigma=4$

$$\text{Shift}=u_2=9.8\Omega_4^{\frac{2}{3}}v^{\frac{1}{3}}n, \quad (4.29)$$

$$\text{Half-width}=\omega_{\frac{1}{2}}=11.4\Omega_4^{\frac{2}{3}}v^{\frac{1}{3}}n. \quad (4.30)$$

In these formulas v is the speed of the perturbing particle (in the center of gravity system).

Equation (4.27) is somewhat embarassing: it makes the shift depend directly upon the cutoff radius and thereby emphasizes the uncertainty of our assumptions. Numerical considerations even show that for electrons as perturbers, (4.27) is generally much greater than (4.28), and this exposes an inconsistency of the present method. Table V shows the ratio $u_2/\omega_{\frac{1}{2}}$ from (4.27) and (4.28) for different temperatures and electron densities.

Burkhardt,[37] Lindholm,[8c,15b] and particularly Unsöld[18] couple an interesting consideration with the analysis leading to Eqs. (4.29) and (4.30), where no convergence difficulties are encountered. If $\sigma=4$, Eqs. (4.22) and (4.23) read

$$u_2=nv\int \sin(\rho_0/\rho)^3\cdot 2\pi\rho d\rho, \quad (4.31)$$

$$\omega_{\frac{1}{2}}=4nv\int \sin^2\tfrac{1}{2}(\rho_0/\rho)^3\cdot 2\pi\rho d\rho, \quad (4.32)$$

because in this instance $\Delta\varphi$, given by (4.26), takes the form $(\rho_0/\rho)^3$ if we put $\rho_0=(\pi\Omega_4/2v)^{\frac{1}{3}}$. This, is the value of ρ (always in terms of our single-impact analysis) for which the phase change $\Delta\varphi=1$ and therefore corresponds to Weisskopf's critical radius ρ_c.

Consider the integrands in (4.31) and (4.32). Clearly $\sin(\rho_0/\rho)^3$ fluctuates rapidly about zero for small ρ and becomes monotone and decreasing for large ρ. The initial fluctuations practically make u_2 depend on the contributions it gets from large ρ. But as for (4.32), its integrand is never negative; hence, the initial fluctuations contribute greatly to $\omega_{\frac{1}{2}}$ while the small contributions from large ρ are less significant. Now the division between large and small ρ comes at $\rho=\rho_0$, the point beyond which neither integrand oscillates.

‖ The Debye-Hückel radius would be a better limit; but the results should not depend critically on this choice if they are to be believed.

TABLE V. Electron broadening. Ratio of shift to half-width in the linear Stark effect.

T(°K)$\backslash n$	10^{12}	10^{13}	10^{14}	10^{15}	10^{16}
10^2	8.6	3.5	2.6	1.4	1.05
10^3	22	11.7	6.3	3.5	1.9
10^4	58	30	16	8.6	3.5
10^5	158	81	42	22	11.7

All this suggests that perhaps a useful simplification results if we integrate (4.31) from ρ_0 to ∞, and (4.32) from 0 to ρ_0. Furthermore, we replace the sine in (4.31) by its argument and revert to the original form for u_2, Eq. (4.20). It reads

$$u_2=\nu_c\sum_j P(\varphi_j)\varphi_j\equiv\nu_c\bar{\varphi}_j. \quad (4.33)$$

The line shift, which arises primarily from small $\Delta\varphi$ (large ρ) is in effect the average number of phase shifts per second.

The integration of (4.32) from 0 to ρ_0 changes the numerical factor in (4.30) very little; hence, we may conclude that, under conditions in which (4.32) is valid, contributions from $\Delta\varphi$ smaller that 1 may be ignored. This, however, is not the same as saying that (4.32) is valid when $\Delta\varphi\gg 1$. Within the limits set by the example of the second-order Stark effect, and for force-law parameters σ greater than 4, one arrives at the qualitative rule that impacts within the optical radius broaden, impacts outside the optical radius shift the line. This somewhat precarious generalization is belied by Eqs. (4.27) and (4.28), and by Table V; hence it has no meaning for first-order Stark effects.

What are the conditions for validity of (4.22) and (4.23) of which (4.31) and (4.32) were special cases? We have derived them by assuming the impacts to be sudden. A little reflection shows that the derivation is essentially unchanged provided the collisions do not overlap, i.e., are binary. The case of multiple collisions has been treated by Lindholm (though in a way which makes no allowance for the *vector* superposition of fields in the Stark-effect problem and which remains within the "classical path" approximation) who succeeded in showing that the foregoing results of this section have wider validity than the present analysis suggests. Some further consideration is given by Krogdahl[39] and, by Kolb.[38]

B. Core and Wing Theorems for Individual Collisions

Under the assumption of binary collisions the criteria for the impact and the statistical theory take special forms.

[39] M. Krogdahl, Astrophys. J. 110, 355 (1949).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

1. Core Theorem

The impact theory (see Sec. I) was established on the assumption that the impacts are sudden and separated in time. The general form for the intensity distribution depends on integrals of the form $\int e^{i[\varphi(t)-\Delta\omega\cdot t]}dt$. A single impact, and therefore the change in φ corresponding to a single impact, are completed in a time of the order of ρ/v. If the change in $\Delta\omega\cdot t$ over this time interval is small compared to unity, i.e., if $\Delta\omega\rho/v\gg 1$ the impact can be considered to be sudden. This inequality can always be satisfied for sufficiently small $\Delta\omega$, hence the impact theory is valid in the core of the line (core theorem).

For *single impacts* the phase change is (see Sec. IIA)

$$\Delta\varphi=\frac{C_\sigma\pi^{\frac{1}{2}}}{\hbar v\rho^{\sigma-1}}\frac{\Gamma[(\sigma-1)/2]}{\Gamma(\sigma/2)}.$$

With the result we can derive[1f] a more specific criterion for validity of the impact theory. If we assume that the important impacts are those that occur within the optical radius ρ_c for which $\Delta\varphi\approx 1$, we have

$$\rho_c\approx(C_\sigma/\hbar v)^{1/\sigma-1}.$$

Combining this equation with the foregoing inequality we have for the range of validity of the impact theory in isolated collisions

$$\Delta\omega\ll v(\hbar v/C_\sigma)^{1/\sigma-1}.$$

In certain cases Kolb[22,38] has shown that this relation is valid under more general conditions than the restrictive ones imposed for this simple derivation.

2. Wing Theorem

The wing theorem, Sec. I, shows that for sufficiently large $\Delta\omega$ (measured from the normal line position) the statistical theory holds, but it does not say beyond what frequency it may be applied. *When the line arises from individual impacts*, a very simple argument covers this point. For there appears in the entire formalism only one time (and hence one frequency) that is characteristic of an impact namely, the duration of the perturbation τ_p. Hence, there is only one frequency τ_p^{-1} to which appeal can be made as a critical limit. The condition, $\Delta\omega\to\infty$, may therefore be replaced by

$$\Delta\omega\cdot\tau_p\gg 1. \qquad (4.34)$$

A more pictorial demonstration of the meaning of this inequality has been given by several authors, among whom Burkhardt[37] seems most explicit and circumspect. He pictures different collisions as producing different rectangles on a perturbation energy *vs* time plot, as drawn in Fig. 5. Distant passages have large τ_p but small $\Delta\omega$, for in this instance we may identify a specific $\Delta\omega$ within the observed line with a specific passage. A given passage is assumed to affect recorded frequencies in two ways, by virtue of the phase jumps occurring at the beginning and the end of a perturbation, and by the actual change of frequency during the perturbation. The former leads to a formula like (4.21) (in which ν_c is strictly *twice* the collision frequency). Empirically, the resulting width is usually small.

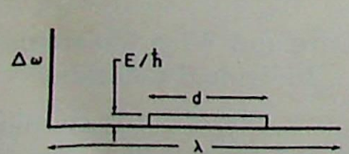


FIG. 5. Idealized representation of collisions. Long and flat rectangles represent gentle impacts of long duration; short and tall rectangle; represent energetic but brief impacts.

The detuning over an interval τ_p requires different treatment in different cases. Taken by itself, a given rectangle in Fig. 5 produces a line at the frequency $\Delta\omega$ and of Lorentz width $1/\tau_p$. Consecutive impacts yield superpositions of such lines, and these superpositions produce a statistical distribution about the different values of $\Delta\omega$ if $1/\tau_p\ll\Delta\omega$, i.e., if the different impact lines do not overlap. Here we discover again the inequality (4.34).

This reasoning is alone perhaps not quite convincing; for it will occur to those with experience in Fourier analysis that the replacement of a set of continuous perturbations by the rectangles of Fig. 5 is a risky practice. Still this demonstration possesses merits, even if rigor is not among them.

There is some likelihood of confusion in the criterion (4.34). We noted, in the text following Eq. (4.32), that the condition $\Delta\varphi\gg 1$ selects those encounters which make important contributions to the *impact* width of the line. This same condition is implied by (4.34). But (4.34) is also the condition for validity of the *statistical* theory. If impact broadening and statistical broadening were mutually exclusive or contradictory effects we should encounter here a paradox. This, however, is not the case. Inequality (4.34) means what it says: it permits the use of statistical theory. The earlier information is also true; it means simply that an impact theory, when carefully employed under these conditions (which is in general more difficult to do), must give approximately the same answer. The literature contains many instances showing this to be true.[8d] Lindholm's theory, for instance, leads to Margenau's statistical formula for Van der Waals broadening, a case which has been repeatedly discussed in this connection.

The overlapping of impact and statistical domains is further illustrated by the following coincidence. We saw that the line shift (4.33) arises from small phase changes, which violate the criteria $\Delta\varphi\gg 1$. But if we compute it under certain circumstances, it is nevertheless the statistical shift. Assume, for instance, that every φ_j arises from a perturbation ΔE_j lasting a time t' which is inversely proportional to the velocity v (so that the interaction distance, d, is constant). Then

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by eGangotri Foundation USA

$\varphi_j=(\Delta E_j/\hbar)(d/v)$. Also, $\nu_c=nqv$ and by (4.33), $u_2=nqd\Delta E/\hbar$, a result which displays the important characteristic of all statistical theories inasmuch as it does not depend on v. In fact, it is exactly the statistical result. In Fig. 6, λ is the mean free path. Along it $\Delta E=0$ except in a region of length d. Hence,

$$\Delta\omega=\frac{\langle\Delta E\rangle}{\hbar}=\frac{1}{\hbar}\frac{0\cdot\lambda+\Delta E\cdot d}{\lambda}=\frac{nqd\Delta E}{\hbar}$$

because $nq\lambda=1$.

C. Some Numerical Estimates, Mainly Regarding the Balmer Lines

The wing theorem for binary collisions says, in effect, that for $\Delta\omega$ greater than ω'', such that $\omega''\tau_p=1$, a statistical description is appropriate. But as we have seen, $\omega''\tau_p$ also equals $\Delta\varphi$ if this simple picture holds.

We now consider the *first-order Stark effect*. Here $\Delta\varphi$ is given by Eq. (4.25). Putting

$$\Delta\varphi=(\pi\Omega_2)/(v\rho_c)=1$$

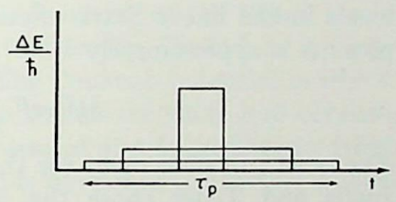


FIG. 6. $\Delta\omega$ *vs* distance travelled by perturber.

we have $\rho_c=\pi\Omega_2/v$. The frequency displacement for this ρ_c is

$$\omega''=\Omega_2/r^2=\Omega_2/\rho_c^2=v^2/(\pi^2\Omega_2). \qquad (4.35)$$

This "statistical frequency limit" within the line is independent of the perturber density.

Let us now consider the half-width of the line which would be calculated by means of impact theory, Eq. (4.28):

$$\omega_{\frac{1}{2}}\approx 2\pi^3\Omega_2^2 n/v. \qquad (4.36)$$

The dependence on v is interesting; the statistical frequency limit moves out farther with increasing v, whereas the impact width becomes smaller.

To make a comparison of (4.35) and (4.36) we need the values of Ω_2. These are discussed by Unsöld (whose C is our $\Omega_2/2\pi$) are are given for the Balmer lines:

	H_α	H_β	H_γ	H_δ	
Ω_2	3.96	10.35	20.5	27.6	(cgs units)

As a typical case we take the H_α line, emitted at a temperature of 10 000°K and an ion density $n=10^{15}$ cm^{-3}. At that temperature v (electron) $=6.23\times10^7$ cm/sec and v (H-ion) $=20.6\times10^5$ cm/sec. From (4.35) and (4.36) one then computes the following values.

TABLE VI. Statistical wavelength limit $\Delta\lambda''$ in Angstroms for the Balmer lines (after Unsöld).

$T=$		25 000°	10 000°	5000°	3000° (K)
H_α	Electrons	580	230	110	70
	Protons	0.63	0.25	0.12	0.08
H_β	Electrons	120	48	24	14
	Protons	0.13	0.05	0.03	0.02
H_γ	Electrons	48	19	9	6
	Protons	0.05	0.02	0.01	0.006
H_δ	Electrons	32	13	6	4
	Protons	0.03	0.01	0.007	0.004

H_α line broadened by

H ions: $\omega_{\frac{1}{2}}=4.8\times10^{11}$ sec^{-1}, $\omega''=1.1\times10^{11}$ sec^{-1}.

electrons: $\omega_{\frac{1}{2}}=1.6\times10^{10}$ sec^{-1}, $\omega''=9.8\times10^{13}$ sec^{-1}.

For all these cases the Holtsmark half width is $\approx10^{12}$ sec^{-1}. The ions, if their broadening effect were computed by the impact theory, would show a half-width greater than ω''. But since for $\Delta\omega\gg\omega''$ the statistical theory is valid one may, so long as one is not interested in the intensity distribution for smaller $\Delta\omega$, disregard $\omega_{\frac{1}{2}}$ for the ions. The electrons, however, produce an ω'' well beyond the limit of interest; hence, a statistical treatment for them is not proper. On the other hand $\omega_{\frac{1}{2}}$ for electrons is very small. The important contribution to the line structure comes from the statistical effects of the ions beyond ω'' and everything else, indeed the entire role of the electrons, can be ignored. The *actual* Holtsmark half width computed for the *ions* under the conditions here assumed is of course greater than ω'' (it is about 8×10^{11} sec^{-1}).

In cases of this sort it is sometimes claimed that the part of the line beyond ω'' can be treated statistically while the inner portion arises from impacts. There is no logical warrant for the second half of this statement despite the core theorem, for the limit of validity of the core theorem need not coincide with that of the wing theorem. Since the interior part has usually been of little interest in line analysis no damage is done by that pleasing supposition.

Table VI, taken from Unsöld,[18] shows the *wave length*, $\Delta\lambda''$, in angstroms from the line center, beyond which the statistical theory is applicable; $\Delta\lambda''=(\lambda^2\omega'')/(2\pi c)$ if λ is the normal frequency of the Balmer line. In all these instances the limit in question is close to the line center for proton broadening, so that Holtsmark's theory may practically be applied throughout the intensity distribution to the protons. The limit for electrons is so far out that they may not be treated statistically. But their impact widths are small, and presumably their entire effect is therefore negligible.

Our logic regarding the electrons contains a blind spot, which we hope the remainder of this article will in part remove. At this point we observe only the proportionality of $\omega_{\frac{1}{2}}$ with n in Eq. (4.36), a feature which

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

indicates that at higher densities the electrons do become important. Had we chosen $n=10^{16}$ cm^{-3} and $T=10\,000$°K in the example of H_α (see text following Eq. (4.36)) $\omega_{\frac{1}{2}}$ for electrons would have been greater than ω'' for protons.

Before concluding we take a glance at the *second-order* Stark effect. Here $\Delta\varphi$ is given by (4.26), and if this is to be 1, $\rho_c=(\pi\Omega_4/2v)^{\frac{1}{3}}$. The corresponding

$$\omega''=\Omega_4/\rho_c{}^4=(2v/\pi)^{\frac{1}{3}}\Omega_4{}^{\frac{1}{3}}.$$

A typical value of Ω_4 is 2×10^{-12} cgs units. Because of the dependence of ω'' on $v^{\frac{1}{3}}$ we encounter here a situation like that presented in Table VI: protons may be treated statistically fairly close to the center of the line, electrons only beyond a large, uninteresting distance from the center. But according to Eq. (4.30) the impact width $\omega_{\frac{1}{2}}=11.4\Omega_4{}^{\frac{1}{3}}v^{\frac{1}{3}}n$. It is proportional to $v^{\frac{1}{3}}$, not to v^{-1} as in the first-order Stark effect. Impact broadening therefore favors the electrons in the present instance.

A numerical example illustrates the situation. If $T=10\,000$°K, $\omega''=7\times10^{11}$ sec^{-1}. Assume again $n=10^{15}$ cm^{-3}. This makes $\Delta\omega_0$ in the Holtsmark theory [$\Delta\omega_0=\Omega_4(F_0{}^2/e^2)$, (compare Sec. III)] equal to 1.36×10^9 sec^{-1}. For the frequency ω'', where this theory becomes valid, the parameter $\omega''/\Delta\omega_0\approx500$, and at this value the function $W[(\omega''/\Delta\omega_0)^{\frac{1}{2}}]$ (see Fig. 2) is already quite small. Hence the Holtsmark theory has little to say about the case. The impact width for electrons, however, leads to $\omega_{\frac{1}{2}}=8\times10^{10}$ sec^{-1} which is the dominant contribution.

The formula for the electron contributed to the line width in the case of the quadratic Stark effect may be written more explicitly. For a hydrogen-like atom (if $Z=1$) or an ion of charge Z with principal quantum number g one finds, approximately,

$$\Omega_4\cong\frac{1}{2}\left(\frac{g}{Z}\right)^6\frac{a_0{}^3e^2}{\hbar}$$

$$=1.4\left(\frac{g}{Z}\right)^6\times10^{-17}.$$

Hence

$$\omega_{\frac{1}{2}}\approx6.6\times10^{-11}\left(\frac{g}{Z}\right)^4nv^{\frac{1}{3}}. \tag{4.37}$$

V. DISAPPEARANCE OF LINES WITH HIGH QUANTUM NUMBERS

Special consideration must be given to the Balmer lines involving high quantum numbers, and in general to lines emitted from highly excited states of atoms or ions within a plasma. The states near the continuum limit have a tendency to disappear for two reasons:

1. Broadening of the levels causes them to merge, which effect begins among the excited states which lie close together on the energy scale. We speak of this as the merging of states.

2. There is also an effective lowering of the continuum limit or, more precisely, an upward displacement of the higher levels because the electron in the atom does not move in a pure Coulomb field. The actual field in a plasma is subject to the Debye-Hückel cutoff. As a result, levels do not approach the energy zero with large quantum number g like $-1/g^2$, but there will be a finite distance between the energy of the level with the largest g (say g^*) and energy zero. This has been recognized in connection with the problem of calculating partition functions, where actually *both* effects, 1 and 2, are important.

Effect 1 has been treated by Inglis and Teller,[40] effect 2 most carefully by Unsöld[41] and by Ecker and Weizel.[42]

The best simple derivation of the formula for the merging of levels (differing only slightly from that of Inglis and Teller) and reliable comments on its precision are found in Unsöld, whose method is followed here. Our analysis will again be applied to the levels of hydrogen, for which $E_g=-e^2/2g^2a_0$ and therefore the undisturbed level separation at large g is

$$\Delta E=e^2/g^3a_0.$$

Here a_0 is the first Bohr radius. The splitting of the levels in the linear Stark effect for the extreme components is approximately

$$\tfrac{3}{2}g^2a_0eF.$$

The most probable value for the ionic field is $4.2\,e^2n^{\frac{2}{3}}$; Inglis and Teller chose the smaller value $3.7\,e^2n^{\frac{2}{3}}$. Merging will take place when the splitting equals $\approx\Delta E/2$. Hence the last discernible level will have the quantum number g^* which is defined by

$$\frac{1}{2}\frac{e^2}{a_0(g^*)^3}=\frac{3}{2}(g^*)^2a_0 3.7e^2n^{\frac{2}{3}}.$$

When logarithms are taken this becomes

$$\log_{10}n=23.3-7.5\log_{10}g^*. \tag{5.1}$$

This formula has often been used for the determination of the ion density n.

So far we have neglected the electrons. If only their statistical effect is considered (assuming that their impact width is small), Eq. (4.28), which represents the limit beyond which the statistical theory holds, tells what role is played by the electrons. For if $\hbar\omega''$ is much greater than the above $\Delta E/2$, they do not contribute to the ionic breadth when it equals $\Delta E/2$ and may therefore be neglected. But if

$$\hbar\omega''=(\hbar v^2)/(\pi^2\Omega_2)<\Delta E/2$$

[40] D. R. Inglis and E. Teller, Astrophys. J. **90**, 439 (1939). [For experimental verification see F. L. Mohler, Astrophys. J. **90**, 429 (1939).]

[41] A. Unsöld, Z. Astrophys. **24**, 355 (1948).

[42] G. Ecker and W. Weizel, Ann. Physik **17**, 126 (1956).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

the electrons must be considered. For large g, the linear Stark constant [defined in Eq. (4.24)] has approximately the value $e^2a_0g^2/\hbar$. The value of v^2 for electrons is taken from

$$\tfrac{3}{2}kT=\tfrac{1}{2}mv^2.$$

Equation (5.1) is therefore subject to modification on account of the electrons when

$$\frac{\hbar}{\pi^2}\frac{3kT/m}{e^2a_0(g^*)^2/\hbar}<\frac{e^2}{2(g^*)^3a_0},$$

or

$$T<\frac{\pi^2me^4}{6\hbar^2kg^*}=\frac{5\times10^5\ ^\circ\mathrm{K}}{g^*}. \tag{5.2}$$

How to include the electrons when this condition holds —and it holds in many cases of physical interest—is, in principle, a difficult problem. The custom seems to be to accord them a role equal to that of the ions and to replace n in Eq. (5.1) by $2n$. Since this adds to $\log_{10}$ only the amount 0.3, and the number 23.3 on the right of (5.1) is hardly certain to within that amount, it is not reasonable to worry about the electrons in this connection.

The second effect, the drowning of the higher levels in the continuum, is treated by Unsöld in a very simple schematic way. The Coulomb potential in which an electron moves never reaches the value zero because another positive ion is situated at a finite distance from it. The electron may slide over into the trough of the other ion, even at a negative total energy, much in the manner in which high atomic levels are depleted by a strong electric field. If the depth of the transfer channel is made to correspond to the mean distance between atom and nearest ion, the highest permitted quantum g^* number is given by

$$(g^*)^2=(Z^{4/3}/6a_0)\left(\frac{4\pi}{3}n\right)^{-\frac{1}{3}}. \tag{5.3}$$

With this simple model, g^* does not depend on the temperature because it ignores the electrons.

The treatment by Ecker and Weizel involves a solution of the Schrödinger equation for an electron moving in a Debye field (or, if the reader is a nuclear physicist, a Yukawa field) of the form

$$V=-(Ze^2/r)\exp(-r/D)+\text{const.} \tag{5.4}$$

The Debye radius is

$$D=[kT/4\pi ne^2(1+Z)]^{\frac{1}{2}}\approx10^{-5}\text{ cm for }\begin{cases}T=10\,000^\circ\mathrm{K}\\ n=10^{16}\text{ cm}^{-3}\end{cases} \tag{5.5}$$

in terms of Z, the charge on an ion. The result obtained is simple and agrees with the plausible expectation that the largest orbit of the cited electron shall be smaller than D:

$$(g^*)^2a_0/Z\leqslant[kT/4\pi ne^2(1+Z)]^{\frac{1}{2}}. \tag{5.6}$$

For $Z=1$ this means

$$g^*\cong3\times10^4(n/T)^{-\frac{1}{4}} \tag{5.7}$$

if g^* is the greatest possible quantum number.

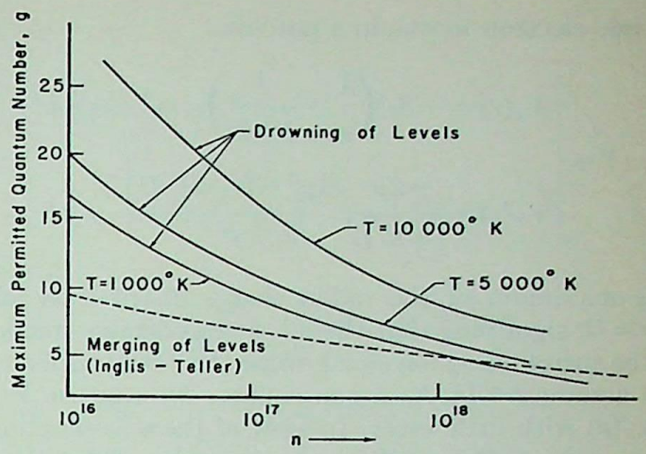


FIG. 7. Disappearance of levels in hydrogen.

Equation (5.7) is plotted for three different temperatures in Fig. 7. On the same figure we have drawn Eq. (5.1). Unsöld's Eq. (3) [Eq. (5.3) of this paper], which is not included, specifies values of g^* which fall slightly below the dotted curve. For high temperatures and low densities the merging of the levels according to Inglis and Teller determines the highest permitted quantum number, whereas in the other extreme the drowning of levels in the continuum is the decisive effect.

The latter has been studied in another way, employing straightforward perturbation theory, in a recent, unpublished calculation.[43] This led to the result

$$(g^*)^2=0.86D, \tag{5.8}$$

which reduces the limiting quantum number slightly below the value given by Eq. (5.6).

Perturbation theory is not strictly applicable since for the last level the perturbation energy is of the same order of magnitude as the unperturbed energy. For this reason we present here the essentials of a parallel variational calculation¶ in which an exponential screening factor $\mathfrak{Z}$ is employed as variable parameter in connection with hydrogen state functions of principal quantum number g. It may be asked why this method should be expected to work, since variation of $\mathfrak{Z}$ might push a hydrogen level initially assigned to the quantum number g down to energies corresponding to lower values of g. The answer is that variation of $\mathfrak{Z}$ does not alter the number of nodes, which is controlled by g, and that therefore a crossing from one g to another cannot occur.

In view of the Debye-Hückel shielding effect the

[43] D. Kelly and H. Margenau, Progress Report, October 1, 1956, Contract Nonr 609 (22).

¶ Contributed by D. Kelly.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

atomic electron moves in a potential

$$V=\begin{cases} V_i(r)=-Ze^2\left(\dfrac{1}{r}-\dfrac{1}{b+D}\right) & r\leqslant b \\ V_0(r)=\dfrac{-Ze^2}{1+b/D}\dfrac{\exp[-(r-b)/D]}{r} & r\geqslant b. \end{cases}$$

The maximum of the radial charge distribution falls at $r=D$, signifying that there is a considerable fraction of the surrounding ion cloud inside the "Debye sphere." We assume $b\ll D$. As a consequence we can use $V(r)=V_0(r)$ with little error. In view of these assumptions concerning b/D and $V(r)$ we write the Schrödinger equation for the optical electron in the form

$$H\psi=-(\hbar^2/2m)\nabla^2\psi-Ze^2(e^{-r/D}/r)\psi=E\psi. \quad (5.9)$$

Choosing functions ψ_g which satisfy the Schrödinger equation for the hydrogen atom of nuclear charge $\mathfrak{Z}$ for quantum number g we compute the integral

$$E_g=\int\psi_g^*H\psi_g d\tau$$

and requires $\partial E_g/\partial\mathfrak{Z}=0$. ($\mathfrak{Z}$ is the variable charge number, Z the fixed ion charge number.) Since

$$-(\hbar^2/2m)\nabla^2\psi_g=[E_g^0+(\mathfrak{Z}e^2/r)]\psi_g$$

and $E_g^0=\mathfrak{Z}^2E_H/g^2$ we have

$$E_g=\frac{\mathfrak{Z}^2}{g^2}E_H-2E_H\int\psi_g^*\left(\frac{\mathfrak{Z}-Ze^{-r/D}}{r/a_0}\right)\psi_g d\tau, \quad (5.10)$$

where $E_H=-e^2/2a_0=-13.53$ ev. Expanding $e^{-r/D}$ in a power series yields

$$\int\left(\frac{\mathfrak{Z}-Ze^{-r/D}}{r/a_0}\right)|\psi_g|^2d\tau$$

$$=(\mathfrak{Z}-Z)\frac{\mathfrak{Z}}{g^2}+\kappa\sum_{m=0}^{\infty}\frac{(-1)^m}{m!}\left\langle\left(\frac{\kappa r}{a_0}\right)^m\right\rangle \quad (5.11)$$

$$\kappa\equiv a_0/D$$

$$\langle(\kappa r/a_0)^m\rangle\equiv\int|\psi_g|^2(\kappa r/a_0)^m d\tau.$$

The values of $\langle(\kappa r/a_0)^m\rangle$ are given by Condon and Shortley.[44] For S states,

$$E_g=E_H\{[(-\mathfrak{Z}^2+2\mathfrak{Z}Z)/g^2]-2Z\kappa S(\beta)\}$$

$$\beta\equiv\kappa g^2/\mathfrak{Z} \quad (5.12)$$

$$S(\beta)\approx\frac{3}{4}\beta-\frac{5}{12}\beta^2+\frac{35\beta^3}{192}-\frac{63}{960}\beta^4+\cdots.$$

The variational constraint, $\partial E_g/\partial\mathfrak{Z}=0$, leads to

$$\mathfrak{Z}-Z+Z\beta^2\partial S(\beta)/\partial\beta=0. \quad (5.13)$$

The principal quantum number of the last level (g^*) is given by the equation $E_g^*=0$. This gives

$$2Z-\mathfrak{Z}-2Z\beta S(\beta)=0. \quad (5.14)$$

Elimination of Z through (5.10) and (5.11) gives an equation involving only β,

$$1-2\beta S(\beta)+\beta^2(\partial S(\beta)/\partial\beta)=0, \quad (5.15)$$

which has $\beta=1.27$ as a root. Thus

$$\beta=[(g^*)^2a_0]/\mathfrak{Z}D=1.27.$$

Reintroducing the true ion charge number, Z, we find

$$(g^*)^2=0.804(ZD/a_0), \quad (5.16)$$

which differs very little from the perturbation result, Eq. (5.8).

We might also inquire how the departure from the hydrogen level structure affects the merging of lines. For this purpose we can apply perturbation theory to all but the last few levels, using hydrogen state functions with $l=l_{\max}=g-1$, since it is for these substates that the Stark splitting is extreme. These functions prove to be very simple and it is not necessary to resort to to the expansion in $\langle(\kappa r/a_0)^m\rangle$. The calculation shows that the first order correction to the hydrogen levels is independent of g and the second order term is so small as to change g^* as given by (5.1) less than one percent.

There are other corrections. Edmonds[45] has calculated, for instance, how the effective mean distance between ions is altered by the Debye shielding effect. For present concerns, relating to the disappearance of the lines, this consideration is of secondary importance.

VI. BROADENING OF DEGENERATE LEVELS

Recently, Kolb[30] has treated the broadening of hydrogen lines in a manner which is more careful than previous treatments in its consideration of the difficulties of degeneracy and which furthermore attempts to include both ions and electrons in their simultaneous actions upon the radiating atom.

We present a synopsis of Kolb's calculations after first discussing certain fundamental ideas of the method and also its principal results. The details given here differ from Kolb's in this respect: our functions [see Eq. (6.8)] are solutions of the *unperturbed* Schrödinger equation, whereas Kolb's are said to be adiabatic functions. The latter choice leads to difficulties which our treatment avoids. Let us recall the meaning of "adiabatic" in the first and second sense discussed in Sec. IIB. The collisions a hydrogen atom can experience with an *electron* can be divided roughly into two extreme groups, namely, (1) *close collisions* that turn the atom around (adiabatic in the second sense) and (2) *distant*

[44] E. U. Condon and G. H. Shortley, *Theory of Atomic Spectra* (Cambridge University Press, London, 1957).

[45] F. N. Edmonds, Astrophys. J. **123**, 95 (1956).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

or weak collisions that do not turn the atom (adiabatic in the first sense). We assume that under certain specifiable conditions the weak collisions are mainly responsible for the broadening, so that the analysis can be restricted to collisions of type 2. One can now speak of adiabatic and diabatic in the following sense; an adiabatic collision leaves the atom in, or returns it to, its original state (permitting the energy to change while the collision is in process), a diabatic collision causes transitions to other final states.

Principal conclusions are (*a*) the electron effects are an important factor in the broadening of the hydrogen lines; (*b*) the adiabatic and diabatic electron effects are of comparable magnitude.

A. Mathematical Foundations

The energy radiated and absorbed per second due to transitions between states m and the state n (providing $E_n^0 > E_m^0$) is given by[10]

$$I_n(\omega) \propto \left\langle \sum_m [\rho_m(0)-\rho_n(0)] \times \lim_{T\to\infty} \frac{1}{T}\left|\int_0^T dt \mu_{nm}{}^c(t) e^{-i\omega t}\right|^2 \right\rangle_{\text{Av}}. \quad (6.1)$$

Here $\rho_m(0)$ and $\rho_n(0)$ are Boltzmann factors for the states having unperturbed energies E_m^0 and E_n^0; $\mu_{nm}{}^c(t)$ is the dipole matrix element defined by

$$\mu_{nm}{}^c(t) \equiv \int \chi_n{}^*(t) \mu \chi_n(t) d\tau, \quad (6.2)$$

where μ is the dipole moment operator. The functions χ_n satisfy

$$i\hbar \dot{\chi}_n(t) = [H_0 + H_1(t)]\chi_n(t), \quad (6.3)$$

provided H_0 is the Hamiltonian for the atom and $H_1(t)$ the interaction between the atom and the perturbing ions and/or electrons. Equation (6.1) is valid if $E_n^0 > E_m^0$. For states (m,n) such that $E_m^0 > E_n^0$, $e^{-i\omega t}$ must be replaced by $e^{i\omega t}$.

Equation (6.1) was derived by calculating the increase in population of the nth state due to the action of the light. Since $E_n^0 > E_m^0$ the first term, $\rho_m(0)\left|\int \mu_{nm}{}^c e^{-i\omega t}\right|^2$ is seen to represent absorption, i.e., transitions from m to n; the second term $\rho_n(0) \times \left|\int \mu_{nm}{}^c(t) e^{-i\omega t}\right|^2$ represents induced emission (electromagnetic field was not quantized) i.e., transitions from $n \to m$. Therefore $I_n(\omega)$ is the net absorption. The average in Eq. (6.1) is to be carried out over different collisions.

Suppose that the states n and m are degenerate in the absence of perturbations. For the level n, we call the collection of states f and label the individual states ϵ_f; for the states m we use i and α_i. Occasionally, we omit the subscripts on ϵ and α. Equation (6.1) thus becomes

$$I_{\epsilon f}(\omega) \propto \left\langle \sum_{i,\alpha_i} [\rho_{\alpha i}(0) - \rho_{\epsilon f}(0)] \frac{1}{2\pi} \times \lim_{T\to\infty} \frac{1}{T}\left|\int_0^T dt \mu_{\epsilon f \alpha i}{}^c(t) e^{-i\omega t}\right|^2 \right\rangle_{\text{Av}}. \quad (6.4)$$

Defining

$$I_{\epsilon f \alpha i} \equiv \frac{1}{2\pi} \lim_{T\to\infty} \left|\int_0^T dt \mu_{\epsilon f \alpha i}{}^c(t) e^{-i\omega t}\right|^2_{\text{Av}}, \quad (6.5)$$

we now sum over all degenerate f states, obtaining

$$\sum_{\epsilon f} I_{\epsilon f}(\omega) \propto \sum_i \{\sum_{\epsilon f \alpha i} [\rho_{\alpha i}(0) - \rho_{\epsilon f}(0)] I_{\epsilon f \alpha i}\}. \quad (6.6)$$

The term in brackets represents transitions between the groups of i and f states. The first term represents absorption, and we define the absorption coefficient I_{if} as

$$I_{if} \equiv \sum_{\epsilon f \alpha i} [\rho_{\alpha i}(0) I_{\epsilon f \alpha i}]. \quad (6.7)$$

In keeping with the weak-collision hypothesis, a solution of Eq. (6.3) is written as

$$\chi_\alpha(t) = \sum_{\alpha'} C_{\alpha'\alpha}(t) \varphi_{\alpha'}{}^0 \times \exp\left[-\frac{i}{\hbar}\int_0^t \{E_{\alpha'}{}^0 + (H_1)_{\alpha'\alpha'}\} dt'\right], \quad (6.8)$$

where $\varphi_{\alpha'}{}^0$ and $E_{\alpha'}{}^0$ are eigenfunctions and energies of H_0, and $(H_1)_{\alpha'\alpha'}$ is given by

$$(H_1)_{\alpha'\alpha'} \equiv \int \varphi_{\alpha'}{}^{0*} H_1 \varphi_{\alpha'}{}^0 d\tau.$$

When Eq. (6.8) is substituted into Eq. (6.3), we obtain

$$\dot{C}_{\alpha''\alpha}(t) = -\frac{i}{\hbar} \sum_{\alpha', \alpha' \neq \alpha''}{}' C_{\alpha'\alpha}(t)(H_1)_{\alpha''\alpha'} \times \exp[-i\{\omega_{\alpha'\alpha''}{}^0 t + P_{\alpha'\alpha''}\}] \quad (6.9)$$

upon using the following abbreviations:

$$\omega_{\alpha'\alpha''}{}^0 = \omega_{\alpha'}{}^0 - \omega_{\alpha''}{}^0; \quad \omega_{\alpha'}{}^0 = E_{\alpha'}{}^0/\hbar$$
$$P_{\alpha'\alpha''} = P_{\alpha'} - P_{\alpha''}; \quad P_{\alpha'} = \frac{1}{\hbar}\int_0^t (H_1)_{\alpha'\alpha'} dt. \quad (6.10)$$

We take as an initial condition that only the state α is present at $t=0$. With weak collisions only the initial state is appreciable at a later time, i.e.,

$$|C_{\alpha\alpha}(0)| \simeq |C_{\alpha\alpha}(t)| \simeq 1; \quad |C_{\alpha'\alpha}(t)| \ll 1 \quad \text{for } \alpha' \neq \alpha.$$

Hence, Eq. (6.9) can be written

$$C_{\alpha''\alpha}(t) = -\frac{i}{\hbar}\int_0^t (H_1)_{\alpha''\alpha} \times \exp[-i(\omega_{\alpha\alpha''}{}^0 t' + P_{\alpha\alpha''})] dt'. \quad (6.11)$$

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

In this approximation Eq. (6.8) takes the form

$$\chi_\alpha(t)=\varphi_\alpha{}^0\exp[-i(\omega_\alpha{}^0t+P_\alpha)]+\sum_{\alpha',\alpha'\neq\alpha}{}'C_{\alpha'\alpha}(t)\varphi_{\alpha'}{}^0\exp[-i(\omega_{\alpha'}{}^0t+P_{\alpha'})]. \quad (6.12)$$

When Eq. (6.12) and a similar one for $\chi_\epsilon(t)$ are used to calculate $\mu_{\epsilon\alpha}{}^c(t)$ the result is

$$\mu_{\epsilon\alpha}{}^c(t)=\mu_{\epsilon\alpha}{}^0\exp[i(\omega_{\epsilon\alpha}{}^0t+P_{\epsilon\alpha})]+\sum_{\epsilon'}{}'\mu_{\epsilon'\alpha}{}^0C_{\epsilon'\epsilon}{}^*(t)\exp[i(\omega_{\epsilon'\alpha}{}^0t+P_{\epsilon'\alpha})]+\sum_{\alpha'}{}'\mu_{\epsilon\alpha'}{}^0C_{\alpha'\alpha}(t)\exp[i(\omega_{\epsilon\alpha'}{}^0t+P_{\epsilon\alpha'})], \quad (6.13)$$

where $\mu_{\epsilon\alpha}{}^0=\int\varphi_\epsilon{}^0\mu\varphi_\alpha{}^0d\tau$ and terms of the order C^2 have been neglected. Equation (6.13) together with Eq. (6.5) determines the absorption coefficient. We are now ready to consider two special cases of interest.

B. Adiabatic Approximation

The off-diagonal elements are neglected in this approximation, only the first term in Eq. (6.13) being used. This simplifies the theory so that it can easily be applied to any two states (α,ϵ). The resulting approximation is general enough to include both electrons and ions and serves as a guide for the more complicated theory below, which includes the off-diagonal elements. On substituting Eq. (6.13) into Eq. (6.5) we have, neglecting the off-diagonal elements,

$$I_{\epsilon\alpha}=\frac{|\mu_{\epsilon\alpha}{}^0|^2}{2\pi}\lim_{T\to\infty}\frac{1}{T}\times\left|\int_0^T dt\exp[-i(\Delta\omega_{\epsilon\alpha}{}^0-P_{\epsilon\alpha})]\right|_{\text{Av}}^2, \quad (6.14)$$

where $\Delta\omega_{\epsilon\alpha}{}^0=\omega-\omega_{\epsilon\alpha}{}^0$. Equation (6.14) can be put into the usual correlation form (see Sec. IV)

$$I_{\epsilon\alpha}=\frac{|\mu_{\epsilon\alpha}{}^0|^2}{\pi}\operatorname{Re}\int_0^\infty d\tau\times\exp(i\Delta\omega_{\epsilon\alpha}{}^0\tau)\langle\exp[-iP_{\epsilon\alpha}(\tau)]\rangle_{\text{Av}}. \quad (6.15)$$

The term $P_{\epsilon\alpha}$ contains interactions involving the *ions* and *electrons*. We therefore write

$$P_{\epsilon\alpha}=P_{\epsilon\alpha}{}^i+P_{\epsilon\alpha}{}^e$$

and assume that the ion and electron averages can be calculated separately. The ions are to be treated in the statistical approximation and therefore permit the interaction to be written as time independent. With this understanding

$$P_{\epsilon\alpha}{}^i=(\Omega_2{}^{\epsilon\alpha}/e)F\tau, \quad (6.16)$$

where F is the instantaneous ion field, e the charge on the ion, and $\Omega_2{}^{\epsilon\alpha}$ are terms connected with the Stark shift of the states (ϵ,α).

The electrons are treated by the phase shift (impact) approximation. Consider

$$\langle\exp[-iP_{\epsilon\alpha}{}^e(\tau)]\rangle_{\text{Av}}=\left\langle\exp\left[-\frac{i}{\hbar}\int_0^\tau\{(H_1{}^e)_{\epsilon\epsilon}-(H_1{}^e)_{\alpha\alpha}\}dt\right]\right\rangle_{\text{Av}}, \quad (6.17)$$

where $H_1{}^e$ is the interaction of the atom with all the electrons in the plasma. In the evaluation of (6.17) we employ Eq. (4.15), following Kolb. This does not require that the impacts are sudden or isolated. It is applicable to scalarly additive perturbations, that is perturbations for which

$$H_1{}^e=\sum_p H_1{}^p. \quad (6.18)$$

The sum here is taken over all the electrons contained in some effective volume. The phase change during the time interval τ produced by a single electron depends on the velocity v, the time of closest approach t_i and on some other parameter l. Since

$$H_1{}^e=\sum_{i,v,l}\mathfrak{N}(i,v,l)H_1(i,v,l), \quad (6.19)$$

where $\mathfrak{N}(i,v,l)$ is the number of electrons characterized by (i,v,l) in the effective volume, $P_{\epsilon\alpha}{}^e$ becomes

$$P_{\epsilon\alpha}{}^e=\hbar^{-1}\sum_{i,v,l}\mathfrak{N}(i,v,l)\int_0^\tau[(H_1(i,v,l))_{\epsilon\epsilon}-(H_1(i,v,l))_{\alpha\alpha}]dt, \quad (6.20)$$

$$P_{\epsilon\alpha}{}^e=\sum_{i,v,l}\mathfrak{N}(i,v,l)\varphi_{i,v,l}(\tau),$$

$$\varphi_{i,v,l}(\tau)\equiv\hbar^{-1}\int_0^\tau[(H_1(i,v,l))_{\epsilon\epsilon}-(H_1(i,v,l))_{\alpha\alpha}]dt. \quad (6.21)$$

Using the results developed in Sec. IV we find

$$\langle\exp[-iP_{\epsilon\alpha}{}^e(\tau)]\rangle_{\text{Av}}=\exp\left[\pi R^2\int dvvn_e(v)[\alpha_v(1)-1]\tau\right], \quad (6.22)$$

$n_e(v)$ being the number of electrons with velocity v per cubic centimeter. The effective volume of the system is assumed to be a sphere of radius R and

$$\alpha_v(1)=\int dlW(l)e^{-i\varphi_{v,l}}. \quad (6.23)$$

Here $\varphi_{v,l}$ is the total phase change produced by an electron of velocity v and parameter l which has its time of closest approach in the time interval τ. Now suppose

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

that the parameter l labels two impact variables ρ and σ, ρ being the distance of closest approach and σ some other characteristic variable defined later. The probability distribution functions for these are

$$W(\rho)d\rho=2\pi\rho d\rho/\pi R^2;\quad W(\sigma)d\sigma=d\sigma/\Sigma, \tag{6.24}$$

where Σ is the total cross section for σ. Then

$$\alpha_v=\int\int W(\rho)W(\sigma)\exp(-i\varphi_{v,\rho,\sigma})d\rho d\sigma, \tag{6.25}$$

and hence

$$\langle\exp[-iP_{\epsilon\alpha}{}^c(\tau)]\rangle_{\text{Av}}=\exp\Big[2\pi\int vn_c(v)dv\int\rho d\rho \times\int(d\sigma/\Sigma)(\exp(-i\varphi_{v,\rho,\sigma})-1)\tau\Big]. \tag{6.26}$$

Finally, we replace $n_e(v)$ by $n_eW(v)$ understanding by n_e the number of electrons per cubic centimeter and by $W(v)$ the Boltzmann distribution function; we write $\varphi_{\epsilon\alpha}(\infty)$ for $\varphi_{v,\rho,\sigma}$, letting (ϵ,α) indicate the two states involved and omitting the (ρ,v,σ) dependence; the symbol (∞) means that the phase change produced by one particle is completed. One then finds

$$\langle\exp[-iP_{\epsilon\alpha}{}^c(\tau)]\rangle_{\text{Av}}=\exp[-(u_1{}^{\epsilon\alpha}+iu_2{}^{\epsilon\alpha})\tau], \tag{6.27}$$

where

$$\begin{pmatrix}-u_1{}^{\epsilon\alpha}\\-u_2{}^{\epsilon\alpha}\end{pmatrix}\equiv\begin{pmatrix}\text{Re}\\\text{Im}\end{pmatrix}2\pi n_e\int_0^\infty vW(v)dv\int_0^R\rho d\rho \times\int_0^\Sigma\frac{d\sigma}{\Sigma}\{\exp[-i\varphi_{\epsilon\alpha}(\infty)]-1\}. \tag{6.28}$$

Substitution of Eqs. (6.27), (6.16) and (6.15′) into (6.15) results in

$$I_{\epsilon\alpha}=\frac{|\mu_{\epsilon\alpha}{}^0|^2}{\pi}\text{Re}\Big\langle\int_0^\infty\exp\Big[i\Big(\Delta\omega_{\epsilon\alpha}{}^0-u_2{}^{\epsilon\alpha}-\frac{\Omega_2{}^{\epsilon\alpha}}{e}F\Big)\tau-u_1{}^{\epsilon\alpha}\tau\Big]d\tau\Big\rangle_{\text{Av(ions)}} \tag{6.29}$$

$$=\frac{|\mu_{\epsilon\alpha}{}^0|^2}{\pi}\Big\langle\frac{u_1{}^{\epsilon\alpha}}{(\Delta\omega_{\epsilon\alpha}{}^0-\Omega_2{}^{\epsilon\alpha}F/e-u_2{}^{\epsilon\alpha})^2+(u_1{}^{\epsilon\alpha})^2}\Big\rangle_{\text{Av(ions)}}. \tag{6.30}$$

We shall return to this equation. At present we wish to develop a formula including the off-diagonal elements, which are taken to be zero in the treatment above.

C. Diabatic Effects in Degenerate Systems

The ions are neglected, so that the atomic states are degenerate, in this approximation. The resulting formulas are useful because they indicate the magnitude of the diabatic effects as far as electron broadening is concerned. In particular, the results of the present development, when applied to the Lyman-α line, and then compared with the adiabatic approximation (when ions are neglected), clearly demonstrate that the adiabatic theory can contain large errors. The method of this section can be generalized to nearly degenerate systems, degeneracy being removed by inclusion of the ions, but the results are difficult to handle and have not been carefully studied.

In the weak collision approximation we assume that $|P_{\epsilon\alpha}{}^c|\ll1$ so that Eq. (6.13) can be written in the form

$$\mu_{\epsilon\alpha}{}^c(t)=\mu_{\epsilon\alpha}{}^0\exp(i\omega_{\epsilon\alpha}{}^0t)\Big\{1+iP_{\epsilon\alpha}{}^e+\sum_{\epsilon'}{}'\frac{\mu_{\epsilon'\alpha}{}^0}{\mu_{\epsilon\alpha}{}^0}C_{\epsilon'\epsilon}{}^*+\sum_{\alpha'}{}'\frac{\mu_{\epsilon\alpha'}{}^0}{\mu_{\epsilon\alpha}{}^0}C_{\alpha'\alpha}\Big\}. \tag{6.31}$$

Since $|P_{\epsilon\alpha}{}^e|\ll1$ and the sum are small, we can write

$$\mu_{\epsilon\alpha}{}^c(t)=\mu_{\epsilon\alpha}{}^0\exp(i\omega_{\epsilon\alpha}{}^0t)\exp(i\Phi_{\epsilon\alpha}), \tag{6.32}$$

with

$$\Phi_{\epsilon\alpha}\equiv P_{\epsilon\alpha}{}^e-i\Big[\sum_{\epsilon'}{}'\frac{\mu_{\epsilon'\alpha}{}^0}{\mu_{\epsilon\alpha}{}^0}C_{\epsilon'\epsilon}{}^*+\sum_{\alpha'}{}'\frac{\mu_{\epsilon\alpha'}{}^0}{\mu_{\epsilon\alpha}{}^0}C_{\alpha'\alpha}\Big]. \tag{6.33}$$

When this is substituted in Eq. (6.5) there results

$$I_{\epsilon\alpha}=\frac{|\mu_{\epsilon\alpha}{}^0|^2}{2\pi}\lim_{T\to\infty}\frac{1}{T}\times\Big|\int_0^T dt\exp(-i\Delta\omega_{\epsilon\alpha}{}^0t)\exp(i\Phi_{\epsilon\alpha})\Big|^2_{\text{Av}}. \tag{6.34}$$

In the weak collision approximation, then, Eq. (6.11) for $C_{\alpha'\alpha}$ becomes

$$C_{\alpha'\alpha}=-\frac{i}{\hbar}\int_0^t(H_1{}^e)_{\alpha'\alpha}\exp(-iP_{\alpha\alpha''}{}^e)dt'\cong-\frac{i}{\hbar}\int_0^t(H_1{}^e)_{\alpha'\alpha}dt'. \tag{6.35}$$

If one assumes, as in the adiabatic case, scalarly additive perturbations [see Eq. (6.19)] then, in view of Eq. (6.35),

$$\Phi_{\epsilon\alpha}\equiv\sum_{i,v,l}\mathfrak{N}(i,v,l)\kappa_{i,v,l},$$

$$\kappa_{i,v,l}(t)=\hbar^{-1}\int_0^t dt'\Big\{\sum_{\epsilon'}\frac{\mu_{\epsilon'\alpha}{}^0}{\mu_{\epsilon\alpha}{}^0}(H_1(i,v,l))_{\epsilon'\epsilon}-\sum_{\alpha'}\frac{\mu_{\epsilon\alpha'}{}^0}{\mu_{\epsilon\alpha}{}^0}(H_1(i,v,l))_{\alpha'\alpha}\Big\}. \tag{6.36}$$

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

Equation (6.34) is formally the same as Eq. (6.14). On the other hand, Eqs. (6.36) are similar to Eq. (6.21). We can therefore use the results of the adiabatic treatment when the proper replacements are made. Our results in that case do not contain the effects of the ions. Instead of Eqs. (6.35) we thus find

$$I_{\epsilon\alpha}=\frac{|\mu_{\epsilon\alpha}{}^0|^2}{\pi}\frac{u_1{}^{\epsilon\alpha}}{(\Delta\omega_{\epsilon\alpha}{}^0-u_2{}^{\epsilon\alpha})^2+(u_1{}^{\epsilon\alpha})^2}, \tag{6.37}$$

$$\binom{-u_1{}^{\epsilon\alpha}}{-u_2{}^{\epsilon\alpha}}\equiv\binom{\mathrm{Re}}{\mathrm{Im}}2\pi n_e\int_0^\infty vW(v)dv\int_0^R\rho d\rho \times\int_0^\Sigma\frac{d\sigma}{\Sigma}[\exp[-i\kappa_{\epsilon\alpha}(\infty)]-1], \tag{6.38}$$

subject again to the use of the phase shift approximation, which accounts for our writing $\kappa_{\epsilon\alpha}(\infty)$ in place of the $\kappa_{i,v,l}$ appearing in Eq. (6.36). The condition $|P_{\epsilon\alpha}|\ll 1$ is violated after a long enough time, and Eq. (6.37) is valid *only* in the wings. In applications, since $|\kappa_{\epsilon\alpha}(\infty)|\ll 1$, we retain only the first two nonvanishing terms in the expansion of $e^{-i\kappa_{\epsilon\alpha}}$. In the wings (6.37) becomes

$$I_{\epsilon\alpha}=u_1{}^{\epsilon\alpha}/[\pi(\Delta\omega_{\epsilon\alpha}{}^0)^2], \tag{6.39}$$

and

$$u_1{}^{\epsilon\alpha}=2\pi n_e\int_0^\infty W(v)vdv\int_0^R\rho d\rho\int_0^\Sigma\frac{d\sigma}{\Sigma}\frac{1}{2\hbar^2} \times\left[\int_{-\infty}^{+\infty}dl\{\sum_{\epsilon'}\mu_{\epsilon'\alpha}{}^0(H_1)_{\epsilon'\epsilon}-\sum_{\alpha'}\mu_{\epsilon\alpha'}{}^0(H_1)_{\alpha'\alpha}\}\right]^2 \tag{6.40}$$

with the understanding that H_1 refers to one electron with specified ρ, v, σ. We shall now apply the formulas of this and the previous section to line broadening by electrons.

D. Electron Broadening in the Adiabatic Approximation

In this section, the adiabatic theory of part B is applied to broadening by electrons only. Here "adiabatic" means that we retain only the first term in Eq. (6.13). This calculation, although restrictive, is useful in estimating (*a*) the error introduced by the electron collisions that fail to satisfy the conditions for an impact theory; (*b*) the effects of close collisions. The value of $\varphi_{i,v,l}(\infty)$, defined by Eq. (6.21) except that the integration is now extended over all time, is (with suitable changes of indexes)

$$\varphi_{\epsilon\alpha}(\infty)=(2\Omega_2{}^{\epsilon\alpha}/\rho v)\cos\theta, \tag{6.41}$$

where

$$\Omega_2{}^{\epsilon\alpha}=\tfrac{3}{2}(e^2a_0/\hbar)\{[n^\epsilon(k_1{}^\epsilon-k_2{}^\epsilon)]-[n^\alpha(k_1{}^\alpha-k_2{}^\alpha)]\}. \tag{6.42}$$

In this expression n^ϵ, $k_1{}^\epsilon$, $k_2{}^\epsilon$ are quantum numbers, in parabolic coordinates, of the state ϵ and θ of Eq. (6.41) is the polar angle describing the location of the position vector of closest approach ϱ in some fixed coordinate system; θ is the σ parameter mentioned earlier. Equation (6.28) then gives

$$u_2{}^{\epsilon\alpha}=0$$

$$u_1{}^{\epsilon\alpha}=2\pi n_e\int_0^\infty vW(v)dv\int_0^{\rho_m}\rho d\rho \times\left[1-\frac{\sin(2\Omega_2{}^{\epsilon\alpha}/\rho v)}{(2\Omega_2{}^{\epsilon\alpha}/\rho v)}\right], \tag{6.43}$$

where ρ_m is a cutoff distance and is chosen as the Debye length of Sec. II. With the definition

$$\delta_m{}^{\epsilon\alpha}\equiv 2\Omega_2{}^{\epsilon\alpha}/\rho_m v,$$

Eq. (6.43) then becomes

$$u_1{}^{\epsilon\alpha}=8\pi n_e(\Omega_2{}^{\epsilon\alpha})^2\int_0^\infty\frac{W(v)}{v}dv\left[0.2094-\tfrac{1}{6}\ln\delta_m{}^{\epsilon\alpha}+\frac{(\delta_m{}^{\epsilon\alpha})^2}{240}+\cdots\right]. \tag{6.44}$$

The velocity average is approximated by replacing the v^{-1} outside of [ ] by its Boltzmann average $\langle v^{-1}\rangle=(4/\pi)(1/\langle v\rangle)$ in Eq. (6.44). We then obtain, finally,

$$u_1{}^{\epsilon\alpha}=(32/\langle v\rangle)n_e(\Omega_2{}^{\epsilon\alpha})^2G(\delta_m{}^{\epsilon\alpha}),\qquad G(x)\equiv(0.2094-\tfrac{1}{6}\ln x+x^2/240\cdots) \tag{6.45}$$

and $\delta_m{}^{\epsilon\alpha}$ is redefined as

$$\delta_m{}^{\epsilon\alpha}=2\Omega_2{}^{\epsilon\alpha}/(\rho_m\langle v\rangle). \tag{6.45a}$$

Let us now return to the two points mentioned at the beginning of this section.

(*a*) According to the criterion for the validity of the phase shift approximation (Sec. IV B), the velocity must satisfy

$$v\gg(\Omega_2{}^{\epsilon\alpha}\Delta\omega)^{\frac{1}{2}}\equiv v_c$$

for the present results to be acceptable. One can estimate the error resulting from extending the velocity integration from zero to infinity by computing the contribution to $\langle v^{-1}\rangle$ from velocities less than v_c. One obtains

$$\frac{\langle v^{-1}\rangle-\langle v^{-1}\rangle_{v<v_c}}{\langle v^{-1}\rangle}=1-\frac{m(\Delta\omega)\Omega_2{}^{\epsilon\alpha}}{4kT}+\cdots. \tag{6.46}$$

For 10 000°K, and for H_β (4861 A) at 30 A from the line center, 90% of $\langle v^{-1}\rangle$ comes from velocities which satisfy the phase shift approximation.

(*b*) It is reasonable to treat close collisions by the Lorentz formula [Eq. (1.2)]. This leads to an $I_{\alpha\epsilon}$ which is similar to Eq. (6.30) (the ion contribution is again neglected) except that $u_1{}^{\epsilon\alpha}$ is replaced by τ_c^{-1}, τ_c being the mean time between collisions. Hence the suggestion that $u_1{}^{\epsilon\alpha}$ can be represented, in an approxi-

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

mate way, as a sum of two terms

$$u_1^{\epsilon\alpha}=(u_1^{\epsilon\alpha})_{\rho>\rho_c}+(u_1^{\epsilon\alpha})_{\rho<\rho_c}, \tag{6.47}$$

where $(u_1^{\epsilon\alpha})_{\rho>\rho_c}$ is the contribution from the weak collision theory, computed according to Eq. (6.32) except that ρ is now allowed to vary from ρ_c to ρ_m. Here ρ_c is the value of ρ which roughly divides the impacts into strong (close) and weak ones, chosen so that

$$\delta_c^{\epsilon\alpha}\equiv 2\Omega_2^{\epsilon\alpha}/(\rho_c\langle v\rangle)=1 \tag{6.48}$$

or

$$\rho_c=2\Omega_2^{\epsilon\alpha}/\langle v\rangle. \tag{6.49}$$

In view of Eq. (6.41), $\delta_c^{\epsilon\alpha}$ is the maximum phase shift that an electron with velocity $\langle v\rangle$ and distance of closest approach ρ_c can produce. But

$$(u_1^{\epsilon\alpha})_{\rho<\rho_c}=\tau_c^{-1}\equiv\pi\rho_c^2 n_e\langle v\rangle=(4\pi/\langle v\rangle)(\Omega_2^{\epsilon\alpha})^2 n_e \tag{6.50}$$

and

$$\begin{aligned}(u_1^{\epsilon\alpha})_{\rho>\rho_c}&=(32n_e/\langle v\rangle)(\Omega_2^{\epsilon\alpha})^2[G(\delta_m^{\epsilon\alpha})-G(1)]\\&=(32n_e/\langle v\rangle)(\Omega_2^{\epsilon\alpha})^2\\&\quad\times[-0.0042-\tfrac{1}{6}\ln\delta_m^{\epsilon\alpha}+\cdots].\end{aligned} \tag{6.51}$$

Comparing (6.50) and (6.51), the contribution due to close collisions can be neglected if

$$\frac{(u_1^{\epsilon\alpha})_{\rho<\rho_c}}{(u_1^{\epsilon\alpha})_{\rho>\rho_c}}\simeq\frac{3}{4}\frac{1}{\ln(1/\delta_m^{\epsilon\alpha})}\ll 1, \tag{6.52}$$

that is, if

$$\ln(1/\delta_m^{\epsilon\alpha})\gg 1.5. \tag{6.53}$$

When (6.53) is satisfied, the quantities expressed by Eqs. (6.51) and (6.45) are approximately equal.

Example: If $n_e=10^{15}$, $T=20\,000$°K and $\Omega_2^{\epsilon\alpha}/A=1.1$ (where $A\equiv\frac{3}{2}e^2a_0/\hbar$) then $\ln(1/\delta_m^{\epsilon\alpha})\approx 6$.

E. Broadening of the Lyman-α Line by Electrons

In this section we treat the broadening of the Lyman lines by electrons only. Both adiabatic and diabatic effects are considered. The wave functions used are the Stark wave functions with the Z axis fixed in space

$$\varphi_+=\frac{1}{\sqrt{2}}(\varphi_{200}+\varphi_{210}),\quad \varphi_-=\frac{1}{\sqrt{2}}(\varphi_{200}-\varphi_{210}),$$

$$\varphi_{21\pm1},\quad \varphi_{100},$$

the subscripts on φ denoting, in spherical coordinates, the quantum numbers nlm.

We use Eqs. (6.39) and (6.40), dropping $\sum_{\alpha'}$ in Eq. (6.40) because of the nondegeneracy of the ground state (100). Letting α designate the state (100), ϵ the states

$$(21\pm1),\quad (+),\quad (-),$$

one finds

$$u_1^{211,100}=u_1^{21-1,100}=0,\quad u_1^{+,100}=u_1^{-,100} \tag{6.54}$$

so that

$$I_{211,100}=I_{21-1,100}=0,$$

$$I_{+,100}=\frac{u_1^{+,100}}{\pi(\Delta\omega_{+,100})^2};\quad I_{-,100}=\frac{u_1^{-,100}}{\pi(\Delta\omega_{-,100})^2}, \tag{6.55}$$

$$u_1^{+,100}=u_1^{-,100}=\tfrac{1}{2}|\mu_{+,100}^{0}|^2\langle 6e^2a_0/v\rho\hbar\rangle,$$

and

$$\left(\frac{6e^2a_0}{\rho v\hbar}\right)\equiv 2\pi n_e\int_0^\infty vW(v)dv\int_\rho^{\rho_m}\rho d\rho\left(\frac{6e^2a_0}{v\rho\hbar}\right). \tag{6.56}$$

The off-diagonal elements contribute twice as much as the diagonal elements to Eq. (6.55). We may therefore write Eq. (6.54) in the form

$$I_{+,100}=\frac{3(u_1^{+,100})_{\rm ad}}{\pi(\Delta\omega_{+,100})^2}, \tag{6.57a}$$

$$I_{-,100}=\frac{3(u_1^{-,100})_{\rm ad}}{\pi(\Delta\omega_{-,100})^2}, \tag{6.57b}$$

where $(u_1^{+,100})_{\rm ad}$ is given by Eq. (6.40) with only the diagonal term, $\mu_{+,100}^{0}(H_1)_{++}$ appearing in the sum. Finally, then,

$$\begin{aligned}(u_1^{+,100})_{\rm ad}&=(u_1^{-,100})\\&=\frac{|\mu_{+,100}^{0}|^2}{3}\frac{144a_0^2e^4}{\hbar^2}\frac{n_e}{\langle v\rangle}\ln\left(\frac{\rho_m}{\rho_c}\right).\end{aligned} \tag{6.58}$$

This result for $(u_1^{+,100})_{\rm ad}$ agrees with the corresponding one obtained from the adiabatic theory of part D of this section, Eq. (6.51) with $\Omega_2^{\epsilon\alpha}/A=2$ and definitions of (6.49), (6.45a), and (6.42). Thus a pure adiabatic treatment (part D) which ignores collision-induced transitions underestimates the broadening by a factor of 3, a conclusion which is important later when an attempt is made to introduce both ions and electrons. Such an attempt succeeds simply only in an adiabatic theory. But with the information now at hand one can estimate the contribution to $u_1^{\epsilon\alpha}$ from the off-diagonal elements (diabatic effects) by using the results of this section.

F. Simultaneous Broadening by Ions and Electrons in the Adiabatic Approximation

In this section both electrons and ions are taken into account in the adiabatic approximation to the theory. The basic formulas are Eqs. (6.27), (6.28), and (6.30). The u_1's are treated in the manner explained in part D. So far as the ions are concerned the averaging process indicated in Eq. (6.30) can be performed with the use of the Holtsmark distribution for the field F since the ions are to be treated in the statistical approximation. Thus, define $W(\Omega)$ to be the probability distribution for

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

the quantity $\Omega \equiv F/e$; Eq. (6.30) then becomes

$$I_{\epsilon\alpha}=\frac{|\mu_{\epsilon\alpha}{}^0|^2}{\pi}\int_0^\infty W(\Omega)d\Omega \times\left\{\frac{u_1{}^{\epsilon\alpha}}{(\Delta\omega_{\epsilon\alpha}{}^0-\Omega_2{}^{\epsilon\alpha}\Omega)^2+(u_1{}^{\epsilon\alpha})^2}\right\} \quad (6.59)$$

and, according to the adiabatic theory for electrons, is $u_2{}^{\epsilon\alpha}$ zero, and $u_1{}^{\epsilon\alpha}$ is independent of Ω. Using the Holtsmark distribution for $W(\Omega)$ and adding together a pair of lines that satisfy $\Omega_2{}^{\epsilon\alpha}=-\Omega_2{}^{\epsilon'\alpha'}$ [this simplifies the results and makes $I(\Delta\omega)$ symmetric] one gets after some mathematical manipulation (see Kolb)

$$I_{\epsilon\alpha}=\frac{2}{\pi}\frac{|\mu_{\epsilon\alpha}{}^0|^2}{(\Omega_2{}^{\epsilon\alpha})^2}\int_0^\infty \xi d\xi \exp[(-4.21\Omega_2{}^{\epsilon\alpha}n_i\xi^3-u_1{}^{\epsilon\alpha}\xi)/\Omega_2{}^{\epsilon\alpha}] \times\left[\Delta\omega_{\epsilon\alpha}{}^0\sin\left(\frac{\Delta\omega_{\epsilon\alpha}{}^0\xi}{\Omega_2{}^{\epsilon\alpha}}\right)+u_1{}^{\epsilon\alpha}\cos\left(\frac{\Delta\omega_{\epsilon\alpha}{}^0\xi}{\Omega_2{}^{\epsilon\alpha}}\right)\right], \quad (6.60)$$

where n_i is the number of ions per cubic centimeter. This equation can be expanded for large and small frequencies ($\Delta\omega_{\epsilon\alpha}{}^0$). We present here only the large frequency expansion, since the other is subject to too many invalidating uncertainties. For convenience we make the following changes in notation:

$$Z_{\epsilon\alpha}\equiv 4.52\Omega_2{}^{\epsilon\alpha}n_i^{\frac{2}{3}}/A, \quad A\equiv \tfrac{3}{2}e^2a_0/\hbar,$$

i.e., $Z_{\epsilon\alpha}$ is the Stark shift (in radians per second) resulting from the mean Holtsmark field strength, and

$$B\equiv\Delta\omega_{\epsilon\alpha}{}^0/Z_{\epsilon\alpha},$$

i.e., B is the shift in units of $Z_{\epsilon\alpha}$ while

$$R\equiv Z_{\epsilon\alpha}/u_1{}^{\epsilon\alpha},$$

i.e., R is a measure of the relative importance of ion and electron broadening. With these changes in notation, for large frequencies ($\Delta\omega_{\epsilon\alpha}{}^0$), Eq. (6.60) becomes

$$I_{\epsilon\alpha}(B)dB=|\mu_{\epsilon\alpha}{}^0|^2dB\frac{2}{\pi}\left\{\frac{1}{R(B^2+1/R^2)}+\sum_{n=2}^{\infty}\frac{(-1)^{n-1}}{(n-1)!} \times\frac{\Gamma[(3n+1)/2]}{B^{(3n-1)/2}}\frac{\cos[\{(3n-1)/2\}\tan^{-1}(BR)]}{[1+1/(BR)^2]^{(3n-1)/4}}\right\}. \quad (6.61)$$

Equation (6.61) reduces to the Holtsmark distribution for ions if we set $u_1{}^{\epsilon\alpha}=0$ ($R=\infty$), for in this limit

$$I_{\epsilon\alpha}(B)dB=|\mu_{\epsilon\alpha}{}^0|^2dB\frac{2}{\pi}\sum_{n=2}^{\infty}\frac{(-1)^{n-1}}{(n-1)!}\frac{\Gamma[(3n+1)/2]}{B^{(3n-1)/2}} \times\cos[\{(3n-1)/4\}\pi]. \quad (6.62)$$

This agrees with Eq. (3.11b) when the coefficients are evaluated. Equation (6.61) reduces to the results of the adiabatic electron theory if $Z_{\epsilon\alpha}=0$.

The electron effects are revealed by comparison of Eqs. (6.61) and (6.62). In addition to the correction factors in each term of the sum in Eq. (6.61), there is a term that depends only on the electrons. In the wings of the line this is dominant since it is of the order B^{-2} whereas the mixed electron ion effects are of the order $B^{-\frac{5}{2}}$. This feature is present in all theories which attempt the fusion of statistical and impact features: the intensity in the statistical theory for large B is proportional to $B^{-\frac{5}{2}}$ that in all impact theories B^{-2}.

Equation (6.61) is applicable to any two pairs of states (ϵ,α), (ϵ',α') for which $\Omega_2{}^{\epsilon\alpha}=\Omega^{2\epsilon'\alpha'}$ since it neglects collision-induced transitions (off-diagonal elements of C). Formally, Kolb does generalize Eq. (6.1) to include the off-diagonal elements. This leads to mathematical difficulties because the $u_1{}^{\epsilon\alpha}$ are then dependent on the ion field strength, which complicates the ion-averaging process. The dependence arises through the $C_{\alpha'\alpha}$ (where $\alpha'\neq\alpha$) (taken as zero in the adiabatic approximation). Under certain conditions, these can be neglected. If, for instance, the average ion field is large the degenerate states are split far enough apart so that the off-diagonal elements vanish. Under other conditions, the term in $C_{\alpha'\alpha}$ that causes the difficulty, namely $\exp(i\omega_{\alpha'\alpha}{}^0t)$ (where $\omega_{\alpha\alpha'}{}^0$ is the separation of two degenerate levels due to the ions) can be replaced by unity. This will be true if the collisions are fast enough, i.e., if the collision time is small compared to $(\omega_{\alpha'\alpha}{}^0)^{-1}$. In this latter case, one can employ the adiabatic theory of Sec. E and estimate the diabatic contribution to $u_1{}^{\epsilon\alpha}$ from a knowledge of the diabatic electron effects, using the procedure given in the treatment of the Lyman-α line. That is, one first determines $u_1{}^{\epsilon\alpha}$ (adiabatic and diabatic) by considering only electrons and then uses this $u_1{}^{\epsilon\alpha}$ in the formula for broadening by ions and electrons in the adiabatic approximation.

As an example of this procedure, Kolb considers the following situation. For $n_e=n_i=10^{16}$ cm^{-3}, $T=15\,000$ °K, and $\Omega_2{}^{\epsilon\alpha}/A=10$. While an analysis of this line to determine the diabatic effects has not been carried out, Kolb estimates that they are of the same order of magnitude as the Lyman-α line. He takes the contribution to the weak collision to be twice the adiabatic contribution Eq. (6.51):

$$(u_1{}^{\epsilon\alpha})_{\rho>\rho_c}\simeq -2\left[\frac{16}{3}\frac{n_e}{\langle v\rangle}(\Omega_2{}^{\epsilon\alpha})^2\ln\delta_m{}^{\epsilon\alpha}\right] \quad (6.63)$$

and for the close collisions

$$(u_1{}^{\epsilon\alpha})_{\rho<\rho_c}=\left[4\pi\frac{n_e}{\langle v\rangle}(\Omega_2{}^{\epsilon\alpha})^2\right] \quad (6.64)$$

assuming that

$$u_1{}^{\epsilon\alpha}=(u_1{}^{\epsilon\alpha})_{\rho<\rho_c}+(u_1{}^{\epsilon\alpha})_{\rho>\rho_c}.$$

The result is shown graphically in Fig. 8.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

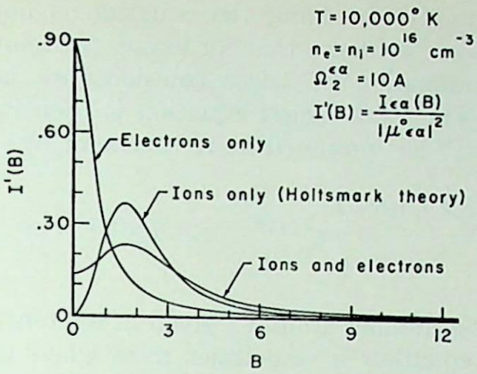


FIG. 8. Line contours produced by different agencies (after Kolb).

This result may be compared with that obtained from the classical theory. Lindholm's result [see Eq. (4.28)] is

$$(u_1^{\epsilon\alpha})_{\mathcal{L}} \equiv \omega_{\frac{1}{2}}/2 = \pi^3(\Omega_2^{\epsilon\alpha})^2(n_e/\langle v\rangle) \times[0.923 - \ln(\pi\Omega_2^{\epsilon\alpha}/\langle v\rangle\rho_m)\cdots].$$

We have replaced the cutoff r_0 by the Debye cutoff ρ_m and v by $\langle v\rangle$. The ratio is then

$$\frac{u_1^{\epsilon\alpha}}{(u_1^{\epsilon\alpha})_{\mathcal{L}}} \approx 0.345\left[\frac{\ln(\rho_m\langle v\rangle/\Omega_2^{\epsilon\alpha})+0.485}{\ln(\rho_m\langle v\rangle/\Omega_2^{\epsilon\alpha})-0.212}\right]. \quad (6.65)$$

For large $\ln(\rho_m\langle v\rangle/\Omega_2^{\epsilon\alpha})$, therefore,

$$u_1^{\epsilon\alpha} \approx 0.345\,(u_1^{\epsilon\alpha})_{\mathcal{L}}. \quad (6.66)$$

The quantity $\ln(\rho_m\langle v\rangle/\Omega)$ is in general large enough so that the difference between Eqs. (6.66) and (6.65) is within the uncertainties of Kolb's method.

For the condition which led to Eqs. (6.63) and (6.64)

$$\ln(\rho_m\langle v\rangle/\Omega_2^{\epsilon\alpha}) \approx 3.25.$$

Equation (6.65) then gives

$$u_1^{\epsilon\alpha} \approx 0.43\,(u_1^{\epsilon\alpha})_{\mathcal{L}}.$$

This is the line width of a single Stark component which is unobservable. The disparity between u_1 and $(u_1)_{\mathcal{L}}$, however, will remain no matter how the components are combined into a resultant contour. Because of the numerous approximations and inherent uncertainties involved in either of the classical path calculations it is difficult to say which formula is numerically preferable, even though Kolb's approach is the more circumspect and realistic one. Later we compare the result of the quantum-mechanical treatment with $(u_1)_{\mathcal{L}}$ (after compounding of Stark states) and find closer agreement.

VII. EXPERIMENTAL RESULTS AND CLASSICAL INTERPRETATIONS

A. Experimental Determination of Plasma Properties

Among the important physical parameters characterizing a plasma are the concentrations of various ions n_i, the electron concentration n_e and the temperature T. In a simple plasma, i.e., one containing but a single type of ion, $n_i = n_e$. The remarks of this section are confined to this case. In complex cases, Dalton's law of total pressure must be employed in addition to the equations discussed below,[46] but otherwise nothing fundamentally new emerges in the analysis of lines from complex plasmas.

A common temperature T exists for all constituents of a plasma if it is in thermal equilibrium. This is not necessarily true in the presence of agencies (discharge currents, ionizing radiations, shocks in rarefied media) which are not part of the thermal mechanism. In such instances it is customary to assign different temperatures to different components (neutral molecules, ions, electrons, photons) and to find ways of measuring each. The limitations of such an approach should, however, be apparent; for the external agencies just mentioned do not always produce a Maxwellian distribution of particle speeds,[47] and a temperature is definable only for a Maxwellian distribution. In a discharge plasma at low electron densities ($n_e < 10^{13}$ cm^{-3}) the temperature concept is known to break down, but it becomes meaningful once more at higher densities when long-range Coulomb forces among the ions and electrons begin to randomize the distribution toward its canonical form. This report does not deal with the pathologies of T, and we assume for the most part that a common T exists.

It has been standard practice to determine $n_i = n_e$ by measuring the intensity distribution of the Balmer lines and comparing it with the Holtsmark formula. Since the temperature does not enter in this procedure, simplicity strongly recommends it. Recent investigations by Edels and Craggs,[48] Lochte-Holtgreven and Nissen,[49] and by Olsen and Huxford[50] are based on it. The risks involved in this method are evident from our earlier theoretical considerations (and from numerous experimental findings of more recent date): reliance on the Holtsmark theory alone is permissible at low ion and electron densities ($n_e < 10^{16}$ cm^{-3}). At high densities, repulsion between perturbing ions produces a smaller intensity in the line wing (where measurements and theory are usually compared), but the electrons add to it. It is likely that a fortunate, rough cancellation between these opposing errors has sometimes led to correct values of n_i even when the method was not strictly valid.

Ion densities obtained by Edels and Craggs and by Lochte-Holtgreven and Nissen in hydrogen arcs were

[46] See, for instance, W. Lochte-Holtgreven, *Temperature, its Measurement and Control in Science and Industry* (Reinhold Publishing Corporation, New York, 1955), Vol. 2. Also Reports on Progress in Physics 21, 312, 1958.

[47] Druyvesteyn, Physica **10**, 61 (1930); **1**, 1003 (1934); H. Margenau, Phys. Rev. **69**, 508 (1946).

[48] H. Edels and J. D. Craggs, Proc. Phys. Soc. (London) **A64**, 575 (1951).

[49] W. Lochte-Holtgreven and W. Nissen, Z. Physik **133**, 124 (1952).

[50] N. H. Olsen and W. S. Huxford, Phys. Rev. **87**, 922 (1952).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

of the order 10^{16} cm^{-3}, in a range where the simple method can hardly fail. Those reported by Olsen and Huxford are somewhat higher (condensed flash discharge) and very probably subject to some error. One should also mention that "fitting a Holtsmark contour" is not a wholly obvious matter. It involves the composition of many Stark-broadened component lines into a single distribution with appropriate weights. The weighting has not always been performed correctly; there are differences, for example, in Olsen and Huxford[50] and Craggs and Hopwood[51] (factor 2 for perpendicular Stark components).

When n_i has been determined, it is a relatively easy matter to find T from the Saha equation (see the following), the relative intensity of different Balmer lines and other known effects involving n_i and T.[46,49] However, since line broadening is strictly also a function of T (mainly through the electron effect), and since we are furthermore interested in *testing* line broadening theories, it behooves us to discuss briefly some available methods for determining n_i and T that are independent of the details of plasma broadening.

1. Saha-Eggert Equation (Mass-Action Law)

This is a relation correcting n_e and T as follows:

$$\frac{n_i n_e}{n_0}=\frac{2Z_i}{Z_0}\frac{(2\pi m_e kT)^{\frac{3}{2}}}{h^3}\exp\left\{-\frac{\chi-\Delta\chi}{kT}\right\}. \qquad (7.1)$$

Z_0 and Z_i are partition functions for neutral atoms and ions, n_0 and n_i their concentrations. For protons, Z_i has the classical form $(2\pi M_p kT)^{\frac{3}{2}}/h^3$; for neutral H atoms, however, Z_0 consists of two factors, the one just written (since $M_H \approx M_p$) and the sum of states $\sum_n g_n \exp(-E_n/kT)$ with weights $g_n=2n^2$ and E_n = hydrogen energy for principal quantum number n. This sum diverges because $E_n \propto n^{-2}$. Physically, however, this divergence is prevented by the drowning of levels discussed in Sec. V. A good practice, established on theoretical and experimental grounds, is to cut off the sum at $n=6$ or 7.

The quantity χ is the ionization potential for the ion, and $\Delta\chi$ is a correction resulting from the drowning of the high levels or, to put it another way, by the lowering of the ionization potential in the presence of ions. It has been computed by Unsöld, Weizel and Ecker, and others (see Sec. V).

Equation (7.1) holds in thermal equilibrium. Strictly speaking, this means that radiative processes as well as corpuscular collision processes, both those generating ions and destroying ions (recombination), must be balanced in detail.[52] If, for instance, radiation escapes partly from the medium, the equation is not exact. There is some evidence that for low n_e photoionization and recombination by triple collisions are not fully effective. The Saha-Eggert equation is then impaired, and Elwert[53] has proposed its replacement by

$$\frac{n_e}{n_0}=\frac{8.4\times10^5\,kT}{n\chi}\frac{}{g}e^{-\chi/kT}, \quad n_e<7\times10^{16}\text{ cm}^{-3}, \qquad (7.2)$$

where g is a number around 2 given in reference 53.

Saha's equation is sometimes used when the temperatures of different plasma components are different. It is then said to define an ionization temperature T_i. The physical meaning of this concept is far from clear; still it may be a useful parameter. But if the energy distribution of the electrons is not Maxwellian (e.g., discharges with $n_e \approx 10^{14}$ cm^{-3}) the Saha equation may be vastly in error. E. Dewan[54] has derived the analog of (7.1) for electron distribution functions given in references 47 and found departures from Eq. (7.1) by large factors.

2. Absolute Intensity of a Line

The temperature alone is involved in the formula for the absolute intensity of the Balmer lines (in emission):

$$I=\frac{2\pi e^2 h f n^2}{m_e\lambda^3}n_0 l\exp\left(-\frac{E_n-E_0}{kT}\right) \qquad (7.3)$$

(λ = wavelength of line, f = oscillator strength, quantum number, n and E_n refer to upper state of line, l is the thickness of the radiating layer, which must be small, i.e., optically thin). If l and n_0 can be determined with precision, this formula is the most direct and reliable means for determining T.

3. Relative Intensity of Different Lines

Two Balmer lines, one originating in the state of principal quantum number n', the other in state n, and of frequencies $\nu_{n'}$ and ν_n, have an intensity ratio

$$\frac{I_n}{I_{n'}}=\frac{n^2\nu_n A_n}{n'^2\nu_{n'}A_{n'}}\exp\left\{\frac{E_{n'}-E_n}{kT}\right\} \qquad (7.4)$$

if all excited states are in thermal equilibrium. A_n is the coefficient of spontaneous emission. Again, in the absence of external agencies, and at high electron-ion concentration even in their presence, this condition is satisfied. Recent measurements in a hydrogen arc discharge by Edels and Craggs,[55] who compared H_α, H_β, and H_γ, have revealed large departures from formula (7.4).

[51] J. D. Craggs and W. Hopwood, Proc. Phys. Soc. (London) A59, 755 (1947).

[52] For a simple discussion see W. Finkelnburg and H. Maecker in *Handbuch der Physik* (Springer-Verlag, Berlin, 1956), Vol. XXII, p. 308.

[53] G. Elwert, Z. Naturforsch. 7a, 432, 703 (1952).

[54] E. Dewan, Dissertation, Yale, 1957 (unpublished).

[55] See reference 48, p. 562.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

4. Intensity of the Balmer Continuum

The intensity of radiation at a wavelength λ below the Balmer continuum limit of 3646 A is given by the formula[56]

$$I_\lambda=\gamma\frac{16\pi^4 m_e e^{10}}{3^3 c^2 h^5 \lambda^2} n_0 l \exp\left\{\frac{E_1-E_2}{kT}-\frac{hc}{\lambda kT}\right\} \times\left[1+\sum_{n=3,4}\binom{2}{n}^3 \exp\left\{\frac{E_2-E_n}{kT}\right\} +\frac{4kT}{Rhc}\exp\left\{\frac{E_2-E_5}{kT}\right\}\right], \quad (7.5)$$

where n_0 is again the concentration of H atoms, l the layer thickness, E_n = energy of nth quantum state $=-Rhc/n^2$, and R = Rydberg constant. The last two terms in [ ] represent contributions of higher continua in the region of the Balmer continuum.

5. Temperature in Shocks

Shocks are an increasingly important agency for producing high temperatures of short duration, yet long enough to emit measurable spectra lines. Temperatures and ion densities obtainable in shocks surpass by far the limitation of most electrical discharges.[57] Formulas for the temperature in shocks may be derived from the Rankine-Hugoniot equations, are given in Turner's dissertation[57] and, in a more elaborate manner, in Courant and Friedrichs.[58]

Of the five methods thus far surveyed only the first presents means for determining n_e; the remainder are independent ways of finding T. The Saha equation, involves both n_e and T; hence it must always be coupled with one or more of the later formulas. Unfortunately, a determination of n_e from Eq. (7.1) is highly inaccurate because $\Delta\chi$ and T appear in an exponent; the error in T resulting from an uncertainty in a given value of n_e would be much less.

A higher degree of confirmation can be given to a Saha n_e if use is made of the Inglis-Teller formula (Sec. V) which involves n_i $(=n_e)$ and only n_i directly. Although it depends on line widths, details do not enter that equation, and the action of electrons affects it in a minor way. But of course it yields only an estimate.

B. Some Recent Experimental Results

The need for modifications of the statistical theory because of the impacts of electrons was recognized by Unsöld and Lochte-Holtgreven at Kiel where, in consequence, a systematic series of experiments were recently performed on the Balmer series by Jürgens,[56] Griem,[59] Dieter-Henkel,[60] and Bogen.[61] High-intensity carbon electrode arcs maintained in a water channel and supporting currents up to 200 amp were used as sources of radiation, the first six series members being observed for various values of temperature and electron density. Although their experiments were done independently, the sum total of work of the Kiel group exhibits a large measure of unity, justifying a discussion of their findings as a whole.

In these experiments, the perturbations arise from protons and electrons acting together in the discharge; these always occur in equal numbers so that $n_e=n_p$ where n_p is the density of protons per cubic centimeter.

In order to allow the application of a theory of broadening to given experimental contours, values of T and n_e must first be determined, for as we have seen in Sec. II, the validity of a particular broadening theory in general depends upon the values of these parameters. Several independent methods for determining the temperature were employed, such as: measurement of the relative intensities of the successive lines of the series emitted from optically thin portions of the discharge, measurement of the degree of inversion of the lines when emitted from optically thick portions of the discharge, and variation of the intensity of the series continuum. The temperatures calculated by these several methods were found (doubtless in part by good fortune) to agree within a few percent, the values ranging from about 5×10^3 to 1.4×10^4 °K. The determination of n_e was made by the single method of matching the observed contours to an intensity distribution. Two distributions were used, the first being the statistical distribution of Holtsmark and the second a distribution formed by smearing the statistical distribution over the Lindholm impact distribution.

The Holtsmark fitting is performed in the far wings of the line where the inequality, Eq. (4.27), holds. Large discrepancies between this distribution and the observed contours were found, especially in the cases of H_α and H_γ, both of which have undisplaced central components in their Stark patterns which contain large fractions of the total intensities of the lines (38% for H_α and 16% for H_γ). The H_β line, having no undisplaced component, lends itself best to a fit by the statistical distribution. An application of the conditions, Eq. (5.2) to the last resolved member of the series shows that the electrons may be neglected as statistically broadening particles at the temperatures calculated. Hence the perturber density determined by the statistical fit is interpreted as that of the protons only.

The dashed curves of Fig. 9 show the contours as found for the first three lines of the series along with their Holtmark wing fits; these curves were taken from

[56] G. Jürgens, Z. Physik **134**, 21 (1952).
[57] See, for instance, Petschek, Rose, Glick, Kane, and Kantrowitz, J. Appl. Phys. **26**, 83 (1955); E. B. Turner, Dissertation, University of Michigan, 1956.
[58] R. Courant and K. O. Friedrichs, *Supersonic Flow and Shock Waves* (Interscience Publishers, Inc., New York, 1948).
[59] H. Griem, Z. Physik **137**, 280 (1954).
[60] W. Dieter-Henkel, Z. Physik **137**, 295 (1954).
[61] P. Bogen, Z. Physik **149**, 62 (1957).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

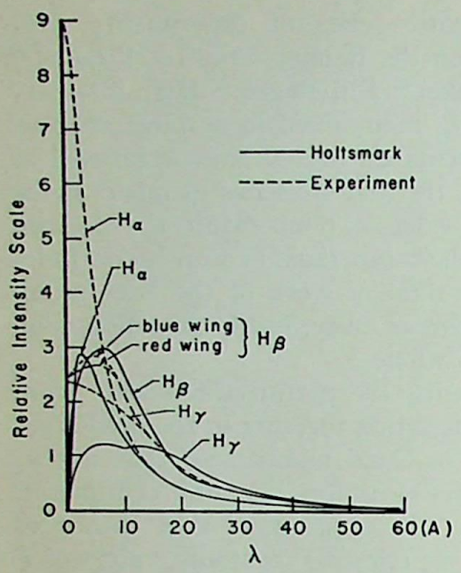


FIG. 9. Comparison of experimental with Holtsmark contours (after Jürgen).

Jürgens and correspond to $T=12\,300$ °K and $n_e=8.4\times10^{16}/\text{cm}^3$.

According to these findings, the experimental contours agree with the statistically calculated ones beyond 30 A, and the agreement occurs for a value of n_e identical with that obtained from the Saha equation. This was disquieting to theorists when the data appeared, for it seemed to show the old but unjustifiable belief in the unimportance of the electrons (at these wavelengths) to be correct. But the later work of P. Bogen[61] has changed this picture.

In his painstaking investigation, which includes a good discussion of errors, Bogen first measures the absolute intensity of the H_β line and uses formula (7.3) to find T. This requires knowledge of n_0, the number of hydrogen atoms, which for the water-stabilized arc employed in these experiments is not readily at hand. It is obtained from a previous study[62] of the properties of this arc, which resulted in the determination of n_0 as a function of T (with use of the Saha equation). Bogen then substitutes T in the Saha equation and obtains $n_e=2.1\times10^{16}$ cm^{-3} for the most fully documented example. When the intensity in the wing of the Balmer lines is plotted one obtains for each of the Balmer lines a graph such as Fig. 10. In the indicated wavelength range the experimental curve lies between two Holtsmark curves, one drawn for $n_e=2.2\times10^{16}$ cm^{-3}, the other for twice that value. For H_α, but not for the higher Balmer lines, the experimental curve approaches the upper Holtsmark curve at about 60 A, and the indication is that the higher Balmer lines likewise show this feature, but the approach takes place at higher wavelengths. Bogen concludes that for large λ the Holtsmark theory is correct for all lines, provided one lets ions and electrons have independent additive effects, i.e., one chooses $2n_e$ instead of n_e as the number of perturbers. The range of correctness coincides with that in which the binary statistical theory (see Sec. III A) is valid. Electron effects appear to be confined to the center of the Balmer lines, and they extend to greater distances from the line center in the higher series members.

We end this review of experimental results with a graphical summary of the shock tube work at Michigan. The latest data of this group may be found in the dissertation of Turner,[57] analyzed by Kolb,[22] from whose thesis Fig. 11 is taken. Again the experimental curve falls outside the Holtsmark distribution for the proper n_e at large distances from the line center. The dashed curve marked "ion and electron theory" represents Kolb's calculation (see Sec. VI), adapted in a suitable way to the Balmer lines, for which a complete calculation is still missing.

C. Classical Interpretations

The electrons must be included in an adequate theory. Their effect, according to all the foregoing developments, is most pronounced in the center of the lines. A simple way to incorporate it is to fold an impact distribution for the electrons into a statistical (Holtsmark) contour. This method, suggested in another connection by Margenau, Burkhardt, and doubtless others, was employed by Griem and Henkel in their data analyses. Griem adopts the use of the distribution:

$$I(\lambda)=\frac{\Delta}{\pi}\frac{I_0(\lambda)}{\lambda^2+\Delta^2}+\frac{\Delta}{\pi F_0}\int_{-\infty}^{+\infty}\frac{\bar{W}(x/F_0)dx}{(\Delta-x)^2+\Delta^2}. \quad (7.6)$$

Here 2Δ is the half-width computed for the electron impacts by the Lindholm theory; $\bar{W}(x/F_0)$ is the statistical distribution for a normal field strength F_0. This is taken from the work of Verwey and Schmaljohann (see Sec. III), which excludes the central undisplaced component of a Balmer line. Hence the central component, whose total intensity is I_0, must be added in via the first term of Eq. (7.1), which shows only an impact width. In Eq. (7.1), λ is measured from the normal position of the line. The bar over $\bar{W}$ is to indicate that a proper average over different Stark components has been taken. Much better representations of the data are obtained through the use of (7.6) as can be seen from the curves of Fig. 12 which are taken from the paper of Henkel.

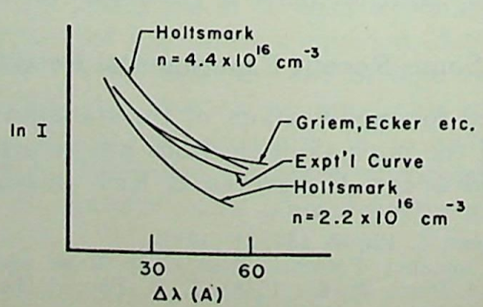


FIG. 10. Comparison of wing intensities (schematic, after Bogen).

[62] Burhorn, Maecker and Peters, Z. Physik 131, 28 (1951).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

A plausibility argument in favor of smearing the distributions takes a form similar to the one used in earlier sections. When a perturbation of the radiator contains two components, one of which acts slowly and the other quickly, the action of the slow one is to produce discrete shifts in energy which may be considered to be constant over time intervals in which many broadening collisions involving the fast component occur. Hence, in our example, each discretely shifted frequency of the statistical theory, produced by the protons, is broadened by the many impacts of the electrons. The condition under which many electron collisions occur during a time in which the ions are essentially motionless is:

$$t_e \ll t_p, \tag{7.7}$$

where t_e is the time between successive electron collisions and t_p is the time during which a proton collision occurs. For classically describable particles we have the definitions,

$$t_e = 1/(n\pi d^2 v_e), \quad t_p = d/v_p, \tag{7.8}$$

where d is defined by Eq. (2.5); v_e and v_p are, respectively, the mean speeds of the electrons and the protons. The inequality (7.2) thus becomes:

$$[(6.90)^3 \pi T^{\frac{3}{2}}/n^{\frac{1}{2}}](m_p/m_e)^{\frac{1}{2}} \gg 1, \tag{7.9}$$

where m_p is the mass of the proton and m_e is the mass of the electrons. This inequality holds for many important applications of electron broadening, both stellar and terrestrial.

It was shown in Sec. II that the classical-path assumption of the impact theory is not necessarily valid for collisions producing phase shifts less than unity. Such phase shifts, however, contribute most of the shift and broadening, as calculated by that theory, for the first order Stark effect [see Eqs. (4.27) and (4.28) and Table V]. Since this is the case, there exists no *a priori* justification for employing the Lindholm distribution, as Griem does in his folding procedure; instead one should replace it with the results of a quantum mechanical calculation. For the quadratic Stark effect this is not the case because the large contributions to the broadening result from collisions falling within the critical radius as has been shown by Unsöld[18] and in Sec. IV of this article.

The peculiarity in the behavior of the first-order Stark effect is connected with the nonconvergence of the broadening when collision parameters out to infinity

TABLE VII. Densities for which $d=\rho_c$ for the first-order Stark effect in hydrogen.

	n^*
H_α	$6.2 \times 10^{18}/\text{cm}^3$
H_β	$1.6 \times 10^{18}/\text{cm}^3$
H_γ	$5.48 \times 10^{17}/\text{cm}^3$
H_δ	$2.52 \times 10^{17}/\text{cm}^3$

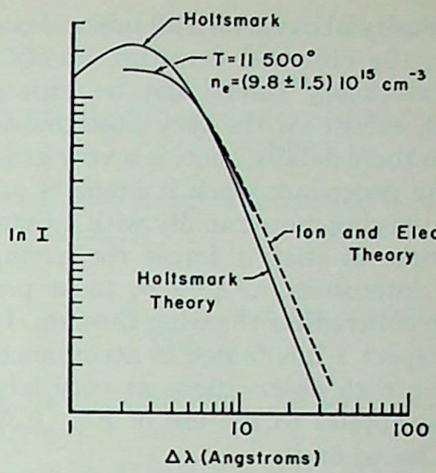


FIG. 11. Comparison of wing intensities (after Kolb).

are considered. The arbitrary cutoff in the collision parameter, taken usually at d or $n^{-\frac{1}{3}}$, introduces an uncertainty in the final broadening; but, more important even, the nonconvergence leads to the phenomenon of the most distant collisions producing the greatest amount of broadening. Griem, in establishing the applicability of the Lindholm distribution to his broadening results, states that only phase shifts greater than unity are important and points to the improbability of multiple impacts within ρ_c, in order to validate the single-impact theory. His arguments, therefore, ignore the long-range contributions to the broadening. In the cases where $d \leqslant \rho_c$, the Lindholm theory may indeed be used. At the temperature of the Kiel experiments $T \approx 10^4$ °K, and we find the values in Table VII

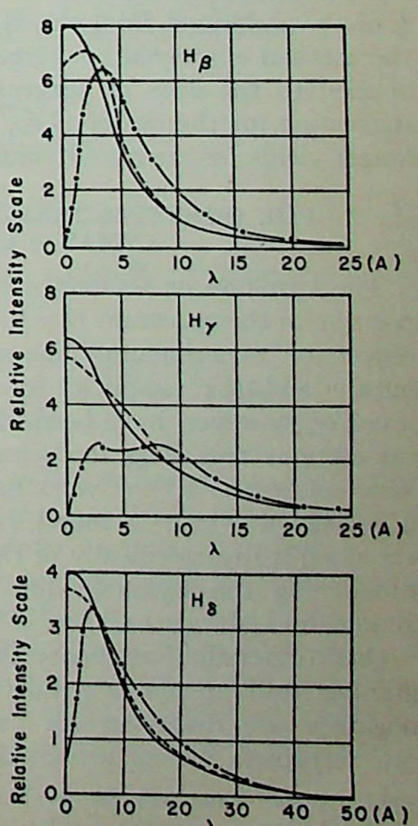


FIG. 12. Comparison of Holtsmark theory jor $F_0=100$ esu, and the folded distribution, with experiment for H_β, H_γ, and H_δ.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

for n^*, the density above which all impacts occur within ρ_c. Thus, for the observed densities, $n \approx 5 \times 10^{16}$ cm^{-3}, the Lindholm theory should not be expected to be strictly valid, except for the very close collisions.

Aside from these details, there is a very general defect of the folding procedure when it involves a statistical distribution varying more rapidly with λ than λ^{-2}, such as Holtsmark's, in that it forces the asymptotic dependence of intensities, as $\lambda \to \infty$, to be proportional to λ^{-2}, which contradicts the wing theorem. Hence, one should not expect a line folded in accordance with Eq. (7.1) to agree with observations at very large λ. This criticism also applies to the use of Eq. (6.30) and all applications based on it.

Ecker[30b] has discussed the desirability of including in the calculation of intensities both the electron effect and the modification of the statistical theory at large distances. He dealt with the latter in reference 30, where the parameter δ is introduced to effect the correction. In reference 30b he folds an impact distribution into a statistical one. But he does this for $\delta = \infty$ and, therefore, obtains results which do not go beyond those of Griem and are in essential agreement with them.

Another result obtained by the Kiel investigators demonstrates that electron collision effects increase at a more rapid rate than do the statistical effects as one proceeds to higher series members. This is in agreement with both the Lindholm theory and the quantum mechanical calculation. Difficulties in the inclusion of the electron effects are implied even in the early work of the Kiel group, as one may see from the following discrepancy. Henkel's fit with the folded distribution leads to values of the perturber density which are about $\frac{1}{3}$ of those derived from the Holtsmark wing fits, i.e., the method employed by Jürgens. When this result is applied to the data of Jürgens, which antedated the suggestion for the use of (7.6), the Saha[59] equation no longer yields his measured value of the temperature.

VIII. QUANTUM TREATMENT OF THE LYMAN-α LINE

The Lyman-α or resonance line of H, because of its position in the spectrum ($\lambda = 1216$ A), has not been the object of experimental line-width work. Its basic nature and the simplicity of the spectral terms it involves, however, have been inviting to theorists, and calculations regarding its behavior have been made. Much of Spitzer's[15a,19,63] work has reference to this line, a review of which is found in Breen's[64] article. An extended study specifically of the effects of electrons in broadening the Lyman-α line was made by Kivel, Bloom, and Margenau.[13]

Quantum radiation theory is applied from the beginning, and numerous details of little practical importance are discussed. In essence, however, three rather typical effects are distinguished. The first is called *universal* broadening. It results from a schematic calculation in which only two levels are included, the lower and the upper states of the line. In the case of L_α, the upper state is actually degenerate, but this degeneracy is not taken into consideration at this stage of the analysis. The results are therefore not specific to the line in question and have a very general meaning. They ignore the first-order Stark effect by not including linear combinations of the degenerate states; and they ignore the second-order Stark effect by neglecting higher energy levels whose virtual excitation normally accounts for the distortion of the radiating atom. What the universal effect describes, then, is broadening by electrons which are scattered nearly elastically, as is apparent in the following.

Polarization of the radiating atom is introduced by properly combining the degenerate states of the upper level. The four correct combinations are known from the theory of the linear Stark effect; they are

$$\psi_+ = \frac{1}{\sqrt{2}}(\psi_{200} + \psi_{210}); \quad \psi_- = \frac{1}{\sqrt{2}}(\psi_{200} - \psi_{210});$$

$$\psi_{21-1}; \quad \psi_{21+1}.$$

The subscripts on ψ denote, in spherical coordinates, the quantum numbers nlm. The first and second of these correspond to different energies $[(-e^2/8a_0) \pm 3ea_0F]$ whereas the last two belong in first approximation to the unchanged energy $-e^2/8a_0$. Moreover, the states ψ_+ and ψ_- have permanent dipole moments in the field F. Transitions between them involve, therefore, spatial *reorientations* of the atomic dipoles occurring as a result of the perturbing electrons. The effect of ψ_{21-1} and ψ_{211} is more subtle; it represents the *induction* of a dipole moment by the passing electron. The calculation which includes only ψ_+ and ψ_- is easy because it does not lead to divergent matrix elements. Its results on the line width were termed "polarization broadening by reorientation." The inclusion of ψ_{21+1} and ψ_{21+1} produces divergences which must be dealt with more carefully. It results in much larger effects, here called "polarization broadening by induction."

Finally, there is an effect connected with matrix elements of the electron perturbation between the states $2p$ and $1s$. These obviously represent quenching by electron collisions, a problem for which the theory had already been given by Wentzel.[65]

Polarization and quenching are strongly dependent on the nature of the quantum levels entering the calculation. The results computed are, therefore, not capable of immediate generalization to other cases and are here only cited as numerical examples. The universal effect, however, presents certain features of wider interest, even though it is small in the case of the Lyman-α line.

[63] L. Spitzer, Phys. Rev. 56, 39 (1939).
[64] R. G. Breene, Geophys. Research Papers No. 41 (1955).
[65] G. Wentzel in *Handbuch der Physik* (see reference 52), Vol. 24/1, p. 738.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

The lengthy calculations in reference 13 concerning the universal effect lead to the same result as a much simpler picture,[17] which we use here. While this analysis standing by itself is perhaps not convincing, its plausibility is fairly high. Instead of considering atom, radiation field, and electron as the quantum system, we concentrate attention on the electron alone and investigate the change in its motion as it meets the radiating atom. If it loses an energy ΔE while it passes the atom, one may say that the atom gains ΔE and therefore emits a frequency shifted to the blue by $\omega_{if}=\hbar^{-1}\Delta E$. What this picture does not suggest is the potential exerted by the radiating atom upon the electron as it moves near it, for the atom has a finite probability of being in the upper as well as the lower state during the emission of radiation.

Here our earlier reflections are helpful. All through Sec. I it was evident that the perturbation responsible for line widths is ϵ, the difference between the upper and the lower energy levels as a function of the position of the perturbing particle. This is the quantity that alters the phase of the radiation in the impact theories and causes the shifts in the statistical theory. Reference 13 bears out the expectation that ϵ, in the form proper for a Coulomb interaction, should be used as the perturbation in the present problem. If the Coulomb interactions between the atomic electron at $\mathbf{r}_R$ and the perturbing electron at $\mathbf{r}$ are denoted by $C(\mathbf{r}_R,\mathbf{r})$, and $C_{ii}(r)$ is the diagonal element of $C(\mathbf{r}_R,\mathbf{r})$ with respect to the ith state of the atom, then

$$\epsilon(r)=C_{22}-C_{11}, \tag{8.1}$$

provided 2 is the upper and 1 the lower state of the line.

In accordance with the usual procedure in (time-dependent) perturbation theory it is now supposed that the electron, before the interaction, has an anergy E_i and its state is represented by a probability amplitude a_i, whereas afterward these quantities are changed to E_f and a_f. We take the states to be those of a free electron: $E_s=(\hbar^2/2m)k_s^2$, $\psi_n\propto\exp(i\mathbf{k}_n\cdot\mathbf{r})$. The amplitude a_f is then a function of t as well as of ω_{if}, and the intensity of the spectral line at a frequency ω_{if} beyond its normal frequency is given by

$$I(\omega_{if})\propto\lim_{t\to\infty}|a_f(\omega_{if})|^2. \tag{8.2}$$

Strictly, the limit $t\to\infty$ is meaningless because the radiative process lasts only for a finite time which is the reciprocal of the natural line width in cycles per second. This natural width is being neglected in the passage to an infinite limit.

The equations for the amplitudes are

$$i\hbar\dot{a}_i=\textstyle\sum_f a_f\epsilon_{fi}e^{-i\omega_{fi}t}, \tag{8.3}$$

$$i\hbar\dot{a}_f=a_i\epsilon_{if}e^{-i\omega_{if}t},$$
$$\omega_{if}\equiv\hbar^{-1}(E_i-E_f). \tag{8.4}$$

If we put $a_i=\exp(-\gamma_u t)$, integrate (8.4) and substitute in (8.3) we find

$$\gamma_u=\sum_f\frac{|\epsilon_{if}|^2}{\hbar^2}\frac{e^{i\omega_{if}t+\gamma_u t}-1}{i\omega_{if}+\gamma_u}. \tag{8.5}$$

The sum over final states is converted to an integral in a well-known manner:

$$\sum_f\to\int\frac{V}{(2\pi)^3}\cdot 2\pi k_f^2dk_f\sin\theta d\theta,$$

where θ is the angle between the wave-number vectors of the electron, $\mathbf{k}_i$ and f_f, and V is the volume per electron. If ω_n is written for $\hbar k_n^2/2m$, the integration over ω_f in (8.5) can be carried out according to the formula[66]

$$\int f(\omega_f)\frac{e^{i(\omega_i-\omega_f)t+\gamma_u t}-1}{i(\omega_i-\omega_f)+\gamma_u}d\omega_f\doteq\pi f(\omega_i),$$

and (8.5) results in

$$\gamma_u=\frac{V}{(2\pi)^2}\frac{m\pi k_i}{\hbar^3}\int|\epsilon_{if}|^2\sin\theta d\theta, \tag{8.6}$$

with the understanding that the remaining integrand is evaluated at $k_f\cong k_i$. This shows that the scattering which leads to γ_u is essentially elastic.

Now ϵ_{if} is easily computed; it is expressible in terms of other atomic integrals frequently encountered in scattering theory, namely,

$$F_{ss}(\kappa)\equiv\int\exp(i\boldsymbol{\kappa}\cdot\mathbf{r}_R)|\psi_s(\mathbf{r}_R)|^2d\mathbf{r}_R.$$

Thus

$$\epsilon_{if}=\frac{4\pi e^2}{V\kappa^2}(F_{22}-F_{11})$$

and

$$\kappa^2=k_i^2+k_f^2-2k_ik_f\cos\theta. \tag{8.7}$$

From this last relation $\sin\theta d\theta$ is computed, and Eq. (8.6) then gives

$$\gamma_u=\frac{2\pi e^4}{V}\frac{a_0^2m}{\hbar^2k_i}\int_0^{x'}\left(\frac{F_{22}-F_{11}}{x}\right)dx. \tag{8.8}$$

The variable x here stands for κ^2a_0 and a_0 is the first Bohr radius; F_{11} and F_{22} refer, of course, to the $1s$ and to the weighted mean of the $2p$ states of hydrogen. The upper limit, $x'=4k_i^2a_0^2$, represents the maximum value permitted by (8.7) when $k_i=k_f$.

Let us call the remaining integral in (8.8) $2\pi^2Y$; it is a function of the initial energy of the perturbing electron through x'. In Fig. 13 it is plotted for the L_α line, E_i being in units equal to the ground-state energy

[66] W. Heitler, *Quantum Theory of Radiation* (Oxford University Press, 1936), first edition, p. 112.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

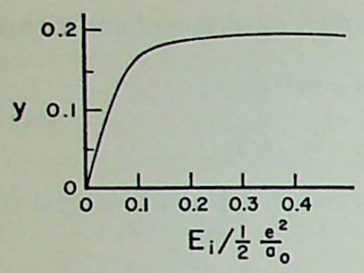


FIG. 13. Graph of the efficiency factor Y *vs* electron energy. See Eq. (8.9).

of hydrogen. If now we replace V in Eq. (8.8) by n^{-1} and $\hbar k_i$ by mv_i, that equation reduces to

$$\gamma_u = nv_i\pi(h/mv_i)^2 Y(v_i). \tag{8.9}$$

From Eq. (8.4)

$$-iha_f = \epsilon_{if}\frac{e^{-\gamma_u t - i\omega_{if}t}-1}{\gamma_u + i\omega_{if}}$$

because $a_i = e^{-\gamma_u t}$. Hence, (8.2) yields a dispersion curve of half-width $2\gamma_u$; from (8.9), therefore

$$\omega_{\frac{1}{2}}(u) = 2n[\pi(h/mv_i)^2]Y(v_i)\cdot v_i. \tag{8.10}$$

A comparison of this result with Lorentz' formula, (1.3), is interesting, for both are of the same form. Evidently the collision cross section for this type of electronic broadening is $\pi\lambda^2$ times an efficiency factor $Y(v_i)$, λ being the DeBroglie wavelength. The rise of $Y(v_i)$ from small values at small v_i illustrates again that a perturbing electron, if it is to produce effects other than shifts described by a static theory, must have sufficient velocity. The development here sketched, and also the work in reference 13, relies upon the Born approximation. When the energy of the impinging electrons is of the order of magnitude of 1 volt, as in the examples under study, this approximation may be greatly in error and the numerical results (e.g., Fig. 13 and Table VIII) may be inaccurate. When good experimental data are at hand, recalculation with better wave functions is desirable.

Meyerott and Margenau[67] computed the analog of Eq. (8.10) with the use of Lindholm's impact theory, basing the work on (4.16). But instead of using a schematic potential of the form (4.17) they employed the potentials ϵ_{2p} and ϵ_{1s} which the electron actually experiences in its approach to the hydrogen atom. The result, plotted in Fig. 14, shows moderate agreement with Eq. (8.10).

TABLE VIII. Broadening of L_α by *electron impacts* at an electron concentration $n=10^{14}$ cm^{-3}, mean energy $(1/25)(e^2/2a_0)$ ($\approx\frac{1}{2}$ ev). The quantity $\omega_{\frac{1}{2}}{}^0$ is the natural line width, 3.12×10^8 sec^{-1}.

Effect, i	$\omega_{1/2}{}^{(i)}/\omega_{1/2}{}^0$
"Universal" broadening	0.13
Polarization by	
reorientation	0.027
induction	2.7
Quenching	0.0013
(Stark broadening by ions	80)

In several publications[68] Sobel'man has shown that Eqs. (8.8) and (8.10) can also be derived in a simple way from scattering theory. His method is quite similar to that employed in reference 67, and he gives useful criteria for the applicability of the scattering approach to line-width problems.

We now turn to the polarization effects, in which electrons cause transitions between the initially degenerate states. Here, the calculation is troubled by the occurrence of spurious divergences in the matrix elements, divergences which can only be avoided by the choice of a cutoff radius like the d encountered in Sec. II. The uncertainties introduced by this somewhat arbitrary procedure are not serious, however. Polarization brings the greatest contribution to the line width so far as the effect of the electrons is concerned.

Kolb, on applying his classical-path theory to the line, has made a careful comparison between his results and those of reference 13. While he reaches numerically different conclusions (resulting from omission of one matrix element in Kivel *et al.*,[13] and also from the fact that Kolb averages the Stark components of L_α with proper weight factors whereas Kivel *et al.* do not average), he is able to show that even the polarization effects can be calculated for the conditions under study with the use of classical-path theory instead of the completely quantum mechanical theory here summarized. The only effect not appearing in Kolb's work is the one called polarization by reorientation, and this is seen to be very small. In Sec. IX it appears that a similar conclusion holds for the Balmer lines—a fortunate fact likely to obviate the need for further detailed calculations.

Finally, there is the width caused by quenching. This, for L_α, turns out to be negligible. In illustration of orders of magnitude the various results are collected in Table VIII, where all quantities refer to a special situation that is approximately realized in the photosphere of the sun. The last line in the table represents the Holtsmark width produced by the ions which in this instance quite evidently outweighs all electron effects.

In judging these results and in appraising the role of the electrons at higher charge densities and for other lines, two facts need to be borne in mind. First, electron effects increase with n, ion effects only with $n^{\frac{1}{4}}$ (see Sec. III). Secondly, lines involving states of higher excitation than L_α offer larger targets to the impinging electrons, and this enhances their sensitivity to broadening impacts. The enhancement is, in fact, greater than the accompanying increase in static polarizability which accounts for the Holtsmark width. While this is not apparent from the present discussion, it manifests itself

[67] R. E. Meyerott and H. Margenau, Phys. Rev. **99**, 1851 (1955).

[68] I. Sobel'man, Optika i Spektroskopiya **1**, 617 (1956); Fortschr. Physik **5**, 175 (1957).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

in our treatment of the Balmer lines in the following section, where the electrons begin to play a role that is no longer subordinate to that of the ions at attainable concentrations. There the work is limited to the effect which was here found to be predominant, namely polarization. In the terminology of Sec. IV, the adiabatic hypothesis is abandoned and full consideration is given to the degenerate states that go into the formation of a broadened line.

IX. QUANTUM THEORY OF THE BALMER LINES

An entirely quantum-mechanical treatment of the broadening of the Balmer lines *due to electrons* has been given by Landwehr.[23] A system composed of the radiator (i.e., an atom or molecule), a number N of perturbing electrons, and the quantized radiation field all confined to a volume V, is described by the Schrödinger equation

$$H\Psi = i\hbar\dot{\Psi}. \tag{9.1}$$

The Hamiltonian H is decomposed into an unperturbed part H^0 representing the total energy of the isolated components of the system, and a part H' which accounts for their mutual interaction. We write

$$H^0 = H_R{}^0 + H_f{}^0 + \sum_{j=1}^{N} H_{pj}{}^0, \tag{9.2}$$

where $H_R{}^0$, $H_{pj}{}^0$, and $H_f{}^0$ are, respectively, the unperturbed Hamiltonians of the radiator, the jth perturbing electron and the radiation field. The perturbation

$$H' = J_{Rf} + \sum_{j=1}^{N} C_{Rj}, \tag{9.3}$$

where J_{Rf} is the interaction between the radiation field and the radiator, and C_{Rj} the Coulomb interaction between the jth electron and the radiator. The Coulomb interactions between the perturbing electrons has been neglected.

Landwehr's method of calculation is similar to that of Weisskoff and Wigner[69] for the problem of the natural line width. It starts with an initial condition in which an excited atom but no photons are present and then derives the probability of finding a photon of a given frequency after a time long enough so that the atom has certainly radiated. At zero time the perturbations are "turned on" suddenly. We then have at $t=0$:

$$\psi(0) = \psi_{n_0}{}^R \psi_0{}^f \psi_{\lambda_0}{}^p,$$

where $\psi_{n_0}{}^R$ is the n_0th eigenstate of $H_R{}^0$, $\psi_0{}^f$ the null state of $H_f{}^0$, and $\psi_{\lambda_0}{}^p$ the λ_0th eigenstate of $\sum H_{pj}{}^0$. The function $\psi_{\lambda_0}{}^p$ is taken to be a product of plane waves. The momentum distribution of the N electrons is assumed to be the equilibrium distribution of the particles at the temperature T (i.e., the Boltzmann distribution).

The solution of (9.1), corresponding to the initial conditions given above, can be expanded over the eigenfunctions of H^0 to give

$$\Psi(t) = \sum_{n\lambda\nu} a_{n\lambda\nu}(t)\psi_n{}^R\psi_\lambda{}^p\psi_\nu{}^f \times \exp\left[-\frac{i}{\hbar}(E_m{}^R + E_\lambda{}^p + E_\nu{}^f)t\right], \tag{9.4}$$

with $a^{(0)}_{n_0,\lambda_0,0} = 1$. The determination of the coefficients a is extremely unwieldy; hence, it is assumed that a satisfactory approximation can be obtained by neglecting all matrix elements of H' that do not involve the initial state. This means that only the initial state can radiate a photon and all collision-induced transitions are from and into the initial state. With this assumption, the equations for the amplitudes become

$$\begin{aligned}
\dot{a}_{n_0,\lambda_0,0}(t) = &-\frac{i}{\hbar}\left\{\sum_{n\lambda}\left(n_0\lambda_0\left|\sum_{j=1}^{N} C_{Rj}\right|n\lambda\right)a_{n,\lambda,0}(t)\right. \\
&\times\exp\left[-\frac{i}{\hbar}(E_n{}^R + E_\lambda{}^p - E_{n_0}{}^R - E_{\lambda_0}{}^p)t\right] \\
&+\sum_{n'\nu}(n_0,0|J_{Rf}|n'\nu)a_{n',\lambda_0,\nu}(t) \\
&\left.\times\exp\left[-\frac{i}{\hbar}(E_{n'}{}^R + E_\nu{}^f - E_{n_0}{}^R)t\right]\right\}, \\
\dot{a}_{n,\lambda,0}(t) = &-\frac{i}{\hbar}\left(n\lambda\left|\sum_{j=1}^{N} C_{Rj}\right|n_0\lambda_0\right)a_{n_0,\lambda_0,0}(t) \\
&\times\exp\left[-\frac{i}{\hbar}(E_{n_0}{}^R + E_{\lambda_0}{}^p - E_n{}^R - E_\lambda{}^p)t\right], \\
\dot{a}_{n,\lambda_0,\nu}(t) = &-\frac{i}{\hbar}(n'\nu|J_{Rf}|n_0 0)a_{n_0,\lambda_0,0}(t) \\
&\times\exp\left[-\frac{i}{\hbar}(E_{n_0}{}^R - E_{n'}{}^R - E_\nu{}^f)t\right].
\end{aligned} \tag{9.5}$$

The approximations have not interfered with the correct normalization, as the solutions of Eq. (9.5) retain the property

$$\sum_{n\lambda\nu} |a_{n\lambda\nu}(t)|^2 = 1.$$

For J_{Rf} we take the single-photon operator

$$J_{Rf} = -(e/m)\mathbf{p}_R \cdot \mathbf{A}_f,$$

where $\mathbf{p}_R$ is the momentum of the radiator's atomic electron and $\mathbf{A}_f$ is the vector potential of the field. The Coulomb interaction, C_{Rf}, is cut off at the distance d

[69] V. Weisskopf and E. Wigner, Z. Physik **63**, 54 (1930).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

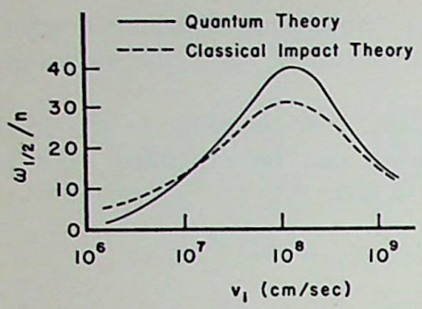


FIG. 14. "Universal" broadening. Graph of half-width *vs* electron velocities.

from the center of the radiator and given by

$$C_{Rj}=\begin{cases}-\dfrac{e^2}{r_j}+\dfrac{e^2}{|\mathbf{r}_j-\mathbf{r}_R|} & \text{for } 0\leqslant r_j\leqslant d\\ 0 & \text{for } r_j<d.\end{cases}$$

It develops that the solution of (9.5) for the initial state amplitude, is of the decaying exponential form when the following conditions hold:

$$\frac{3}{2}\frac{kT}{\hbar}\gg\Gamma_C+\Gamma_J,\quad \omega'\gg\Gamma_C+\Gamma_J.$$

Here ω' is again the frequency of the undisplaced line and $\Gamma_C+\Gamma_J$ is the total decay constant of the initial state, a sum of the contribution from electron scattering Γ_C and from spontaneous emission Γ_J. The inequalities above are both satisfied over wide ranges of interest; the case of electron broadening of the Balmer lines will be analyzed with respect to them below.

There results, for the distribution of amplitudes corresponding to frequency ω due to transitions from an initial state $(n_0,\lambda_0,0)$ to a final state (n',λ_0,ω), the simple formula

$$|a_{n'\lambda_0\omega}(\infty)|^2=\frac{(\Gamma_C+\Gamma_J)/\pi}{(\omega-\omega')^2+(\Gamma_C+\Gamma_J)^2}. \qquad (9.6)$$

Here $\Gamma_J=\sum_{n''}\Gamma^J_{n_0,n''}$ is the total radiative decay constant of the upper state, the sum being taken over all states into which the atom can radiate; $\Gamma_C=\sum_{n'''}\Gamma^C_{n_0,n'''}$ is the total collisional decay constant of the upper state, and the sum is over all states to which the atom can pass by collisional transitions. The Γ_C depends on the transition matrix elements,

$$\Gamma_C\propto F(\kappa)\sum_{n'''}|[n_0|1-\exp(i\boldsymbol{\kappa}\cdot\mathbf{r}_R)|n'']|^2,$$

where $F(\kappa)$ is a function of κ, the wave-number difference between the initial and final states of the colliding electron ($\boldsymbol{\kappa}=\boldsymbol{\kappa}_i-\boldsymbol{\kappa}_f$), and r_R is the coordinate of the atomic electron.

When degeneracy exists in the states of the radiator, the largest contribution to Γ_C comes from matrix elements of the states that are degenerate with the initial state. In addition, $F(\kappa)$ is highly weighted for values of κ which correspond to elastic scattering. The evaluation of Γ_C is, therefore, restricted to those states that are degenerate with the initial state. These matrix elements may be summed approximately in a convenient manner by using known sum rules.

The method leads to a spectral contour associated with a particular radiative transition between given initial and final states. In order to form a single spectral line, such as a Balmer line of hydrogen, it is necessary to combine the results obtained for each transition from one of the initial degenerate states with one of the final degenerate states. In this procedure each of the initial states is assumed to be equally populated and each transition is weighted by the relative transition probabilities per unit time in the dipole approximation. The line formed in this manner is symmetric about ω' but is not of strict resonance shape, being lower in the line core and higher in the wings.

The diagonal matrix elements of H' have been neglected in Eq. (9.6). These represent shifts in the individual line components but are presumably small because of the uniformity of the perturber's charge distribution. These neglected matrix elements, of course, are responsible for the line width and shifts in a purely adiabatic theory; however, in the extreme diabatic case (collision-induced transitions) they contribute only to the shift.

Landwehr has applied this theory to the first three Balmer lines. The half-width Γ_C arising from collision broadening of the atomic state $(n_0l_0m_0)$ is expressible in the form

$$\Gamma^C_{n_0l_0m_0}=n_0^2\left(\frac{2\pi m}{kT}\right)^{\frac{1}{2}}e^4n_ea_0^2\left\{A-B\ln\left(\frac{\hbar^2}{n_e^{\frac{1}{3}}2mkTn_0a_0}\right)+C\mathfrak{E}(\hbar^2/2mkTn_0^2a_0^2)\right\}, \qquad (9.7)$$

where A, B, and C are constants dependent on $(n_0l_0m_0)$ and $\mathfrak{E}(x)$ is the exponential integral of x.

Table IX, calculated for $n_e=5\times10^{16}$ cm^{-3} and $T=10^4$ °K, lists the ratio $\Gamma^C_{n_0l_0m_0}/\Gamma^C_{n_022}$ for the first three Balmer lines.

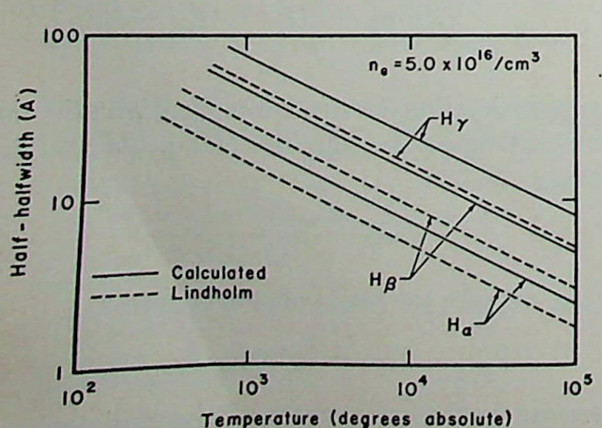


FIG. 15. Broadening of H_α, H_β, and H_γ as calculated by Landwehr, compared with classical impact theory.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

The form of Eq. (9.7) is strikingly similar to Eq. (4.28) for the impact half-width. Indeed when comparison is made of the results of the two theories we find exactly the same variation with temperature and electron density, as shown in Fig. 15. The dotted line is an application of the Lindholm theory, as adapted to the Balmer lines by Griem.[59] Griem treats the normally degenerate states in a given excited level as one state using an average splitting factor for this composite state. The quantum mechanical values of the half-width of the lines exceed those of the classical impact theory, by hardly more than the possible error introduced by Landwehr's use of sum rules in the evaluation of matrix elements. Thus the two theories yield about the same results, the major difference being that the Lindholm theory produces a resonance contour and the quantum mechanical theory a less peaked distribution.

The form of the variation of the collision damping constant with T and n_e

$$\Gamma^C \propto n_e T^{-\frac{1}{2}}$$

is a quite general result and has been found also by Rudkjobing[70] in a quantum mechanical calculation of a different sort.

TABLE IX. $\Gamma^C_{n_0 l_0 m_0}/\Gamma^C_{n_0 22}$ for $n_e = 5\times10^{16}$ cm^{-3} and $T = 10^4$ °K.

		(l_0, m_0)								
		(2,2)	(2, −2)	(2,1)	(2, −1)	(2,0)	(1,1)	(1, −1)	(1,0)	(0,0)
n_0	3	1	1	1.7	1.7	3.4	1.9	1.9	5.5	3.6
	4	1	1	2.8	2.8	2.9	1.3	1.3	4.7	2.8
	5	1	1	3.0	3.0	3.7	1.5	1.5	4.1	2.4

Figure 16 shows the complete quantum contour of H_β taken at $n_e = 5\times10^{16}$ cm^{-3} and $T = 1.25\times10^4$ °K as compared with a resonance curve (as predicted by Lindholm theory) of the same half-width and total integrated intensity. The solid curve in this figure results from a superposition of curves based on Eq. (9.6) for the different degenerate upper state.

Figure 17 compares the values of the half-width of H_β produced by electrons with that produced by ions (Holtsmark theory) where the average broadening is given by[59]

$$\Gamma_H \approx (3\hbar/\pi m)7.8n^{\frac{2}{3}}.$$

For stellar applications at $n \leqslant 10^{15}$ cm^{-3}, $T > 5\times10^3$ °K the electrons do not contribute much broadening of the line and may be omitted from consideration. In some interesting terrestrial applications, such as those discussed in Sec. VII B, the electrons do contribute large effects to the line. Electron broadening becomes inappreciable for extremely large values of the temperature if the number density of electrons is held fixed.

[70] M. Rudkjobing, Ann. astrophys. 12, 229 (1949).

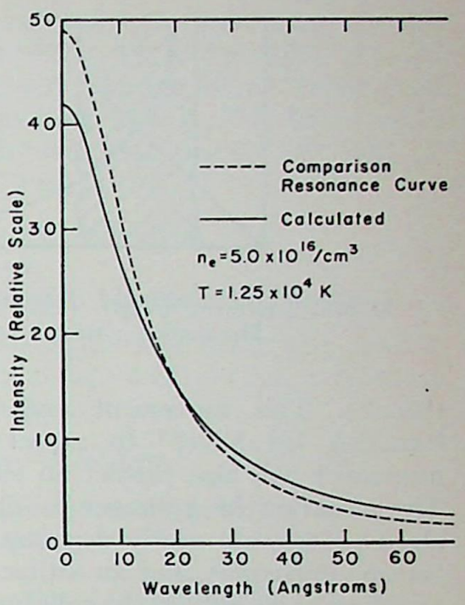


FIG. 16. Comparison of line shapes. Landwehr calculation against resonance curve.

X. SHIFTS PRODUCED BY ELECTRON COLLISIONS**

Argon spectra from high-temperature plasmas produced in shock tubes by Petschek, Rose, Glick, Kane, and Kantrowitz at Cornell proved surprising at first because the lines were not at their expected wavelengths. It seemed that the shifts were greater than could be accounted for by the usual Holtzmark theory of static ion fields and the consequent Stark shifts.†† Hence the results were attributed to line shifts induced by the free electrons in the plasma. Baranger[71] made a quantitative calculation based on the theory of Lindholm (see Sec. IV) which showed that this interpretation was reasonable.

At the same time the group at Yale[13] had been working out the details of the quantum mechanical theory of electron effects. Their predictions of hydrogen line broadening have been shown by Meyerott and Margenau[67] to agree with semiclassical broadening

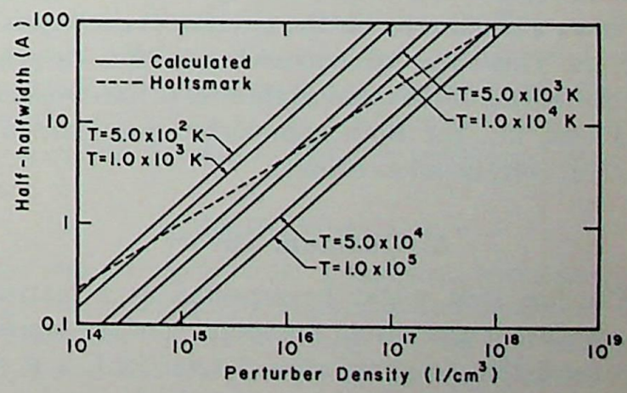


FIG. 17. Broadening of H_β, as calculated quantum mechanically, compared with Holtsmark theory.

** We are greatly indebted to Dr. B. Kivel for a draft of this section.

†† This point has not been entirely clarified; for the asymmetry of the statistical background can introduce a shift.

[71] M. Baranger, Phys. Rev. 91, 436 (1953).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

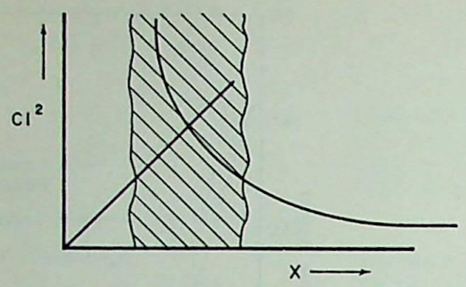


FIG. 18. Matrix element $|C(2p0 \rightarrow 2s0)|^2$ as a function of $x \equiv a_0^2K^2$. The shaded portion is uncertain.

theories. This agreement and the calculation by Baranger led Kivel[72] to expect that the quantum mechanics will also predict an electron-induced shift. The situation in hydrogen is obscured because the atomic levels are nearly degenerate, distant collisions become important, and an artifact such as the Debye shielding to cut off distant collisions must be employed in the calculations. On the other hand, argon presents the difficulty of unknown matrix elements. Helium, having hydrogenlike atomic wave functions and nondegenerate eigenvalues is free of these two problems. The experiment in pure helium, however, is not easy because for elements of low molecular weight it is difficult to heat gas to 20 000 °K and obtain the high free electron densities (10^{17}/cm^3) required. Nevertheless, Seay and Seely[73] of Los Alamos have largely succeeded in carrying it out by using a high explosive driver for their shock tube. The preliminary results are in agreement with the theory when an algebraic error in reference 72 is corrected.

An interesting aspect of this work is that the Born approximation used in the theory gives a reasonable result. Though perhaps somewhat surprising, this reflects the fact that the approximation requires only that the energy of the electron be large compared to the atomic energy separation, and not the atomic binding energies. Also, while the Born approximation fails to some extent for slow electron scattering, in our problem a small error in the electron trajectory is not serious. This same circumstance accounts in general for the success of the classical theories in line-broadening problems, i.e., the Born approximation is close to the classical correspondence limit.

A. Quantum Theory

The line shift might be expected as a matter of course since it arises from the coupling of two states by a perturbation (electron-atom interaction), and such perturbations generally cause coupled states to separate in energy. To deal with this shift quantitatively, we apply a simple correction to the theory of the line width in Sec. VIII where we used the approximation

$$\int f(\omega_f)\frac{e^{i(\omega_i-\omega_f)t+\gamma t}-1}{i(\omega_i-\omega_f)+\gamma}d\omega_f \doteq \pi f(\omega_i). \qquad (10.1)$$

In this formula $f(\omega_f)$ is the collision matrix element integrated over the scattering angle; it depends on $\omega_f \equiv (h/2m)k_f^2$ through the quantity κ, defined in (8.7), which appears in ϵ_{if} and therefore in $f(\omega_f)$. As an example, we reproduce the sketch of $|C(2p0 \rightarrow 2s0)|^2$, which is proportional to $f(\omega_f)$ as a function of the momentum transfer $a_0^2\kappa^2 \equiv x$ given in reference 13 (see Fig. 18). The region of interest in helium is that of large momentum transfers, since the atomic states are not degenerate. Here the close collisions matter, the Debye cutoff does not enter, and the $|C|^2$ dependence on x is known. For hydrogen, however, where the energy levels are close together and the distant collisions become important, the problem is more complicated and is not worked out at this time.

The physical meaning of the decrease in the matrix element with x is that the electron prefers to minimize its momentum change. Consequently, if the electron causes an atomic transition which requires energy (excites the atom) some of this energy will come from the photon (red shift) since the electron is stingy with its own energy change. Similarly for de-exciting collisions the photon will gain energy (blue shift). This is in agreement with simple expectation based on first order perturbation theory where coupled levels repel each other. Figure 19 shows this schematically: the perturbation C changes the optical transitions J in the indicated manner.

Now we return to Eq. (10.1). In view of the slow variation of $f(\omega)$ with ω, Eq. (10.1) is not strictly valid. A more careful evaluation yields an extra, imaginary term on the right, and this term gives rise to a shift, δ_i, of the line.

Quantitatively,[72] the shift (assumed small compared to energy separation; hence the formula given is not valid for hydrogen) is a sum over all coupled states

$$\hbar\delta_i = (hnv_\lambda\sigma_\lambda/3)\sum_{n\neq i}(\omega_{ni}/|\omega_{ni}|)(|\mathbf{r}_{ni}|/a_0)^2, \qquad (10.2)$$

where n=electron density, v_λ=electron velocity, $\sigma_\lambda = \pi(\hbar/mv_\lambda)^2$, $|r_{ni}|^2 = |x_{ni}|^2 + |y_{ni}|^2 + |z_{ni}|^2$, ψ_n=free atom function for eigenvalue E_n with space vector $\mathbf{r}$, and $\hbar\omega_{ni} = E_n - E_i$.

The error in Bethe[74] was the use of $|z_{ni}|^2$ instead of the correct $|\mathbf{r}_{ni}|^2$, as given above. It is interesting that the shift does not depend on the magnitude of the energy separation if the electron has enough energy to excite the transition. For a Maxwell-Boltzmann dis-

[72] B. Kivel, Phys. Rev. 98, 1055 (1955).
[73] G. E. Seay and L. B. Seely, Jr., Bull. Am. Phys. Soc. 1, 227 (1956), Washington Meeting, 1956.
[74] H. Bethe in *Handbuch der Physik* (see reference 52), Vol. 24/1, p. 433.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

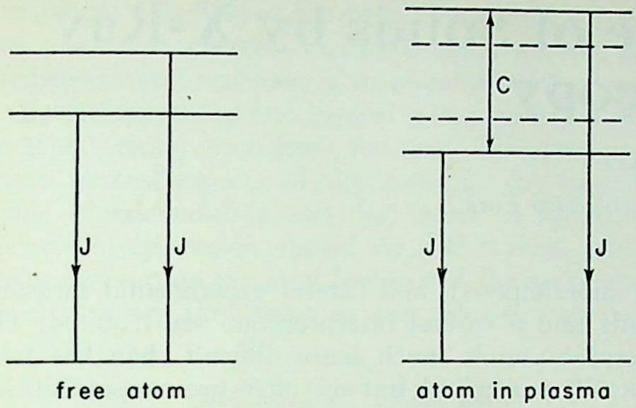


FIG. 19. Energy levels in a plasma.

tribution of velocities we find (δ is in units of sec^{-1}):

$$h\delta_i = 4.55\times10^{-8} hn\left(\frac{e^2}{2a_0kT}\right)^{\frac{1}{2}}\left\{\sum_{E_n>E_i}\left|\frac{\mathbf{r}_{ni}}{a_0}\right|^2 \times\exp(-\hbar\omega_{ni}/kT)-\sum_{E_n<E_i}\left|\frac{\mathbf{r}_{ni}}{a_0}\right|^2\right\}. \quad (10.3)$$

This assumes the Born approximation even for the zero energy electrons.

B. Application to Helium

As a sample application we consider the He line $4s \rightarrow 2p$ at $\lambda = 4713$ A. This is one of the lines studied by Seay and Seely at Los Alamos and has a measured shift to the red estimated to be 25 A when the electron density is $n = 4\times10^{17}/\text{cm}^3$ and the temperature is $T = 20\,000$ °K. This shift is far too large to be accounted for by ions, which produce only a few angstroms. On the other hand, Eq. (10.3) predicts a line shift of about 35 A, as shown below. Considering that the measurement is preliminary and that the calculation uses the Born approximation, this seems to establish the electron effect.

To evaluate Eq. (10.3) we note that the sum over the magnetic quantum number removes the angular dependence[74]; i.e.,

$$\sum_{m'}|\mathbf{r}_{nlm}{}^{n'l+1\,m'}|^2=\frac{l+1}{2l+1}(R_{nl}{}^{n'l+1})^2$$

and

$$\sum_{m'}|\mathbf{r}_{nlm}{}^{n'l-1m'}|^2=\frac{l}{2l+1}(R_{nl}{}^{n'l-1})^2.$$

The radial integrals R are tabulated for hydrogenic wave functions in the quantum mechanical literature.

For the $4s \rightarrow 2p$ transition the most important term in the curly brace is the $4s \rightarrow 4p$ term for which $(R_{4s}{}^{4p})^2 = 540$, $(l+1)/(2l+1) = 1$ and the exponential in Eq. (10.3) is 0.9360, so that there is a contribution of $+505$ (to the red); similarly the transition $4s \rightarrow 5p$ contributes $+56$; $4s \rightarrow 6p$, $+8.4$, $2p \rightarrow 3d$, -4 (to blue); and $2p \rightarrow 2s$, $+9$. The net of all terms with principal quantum number $\leqslant 8$ is 573. The corresponding wavelength shift for $n = 4\times10^{17}$ and $T = 20\,000$ °K ($kT = 1.72$ ev) is

$$\Delta\lambda = \lambda^2\delta/2\pi c = 35 \text{ A}.$$

C. Comparison with Lindholm's Theory

It is interesting to compare this result with that of Lindholm's theory. From Eq. (4.22) we find for the line shift according to Lindholm:

$$u_2 = 9.8\Omega_4{}^{\frac{1}{3}}v^{\frac{1}{3}}n.$$

Considering only the main contributions of $4s-4p$ to the polarization, one finds

$$u_2=\pi^2 nva_0{}^2\left[\frac{|z_{4s,4p}|^2}{a_0{}^2}\frac{e^2/a_0}{hc\Delta\nu}\frac{\alpha c}{v}\right]^{\frac{3}{2}}. \quad (10.4)$$

Since $\alpha = 1/137$, $|z_{4s,4p}|^2 = 180a_0{}^2$, $\Delta\nu_{4s,4p} = 919$ cm^{-1} the line shift for the same values of n and T as in B is about 24 A. We have written (10.4) so that it can be compared with the quantum result, Eq. (10.3), which reads, with retention of the same matrix element,

$$\delta=\pi nva_0{}^2\left[2\frac{|z_{4s,4p}|^2}{a_0{}^2}\left(\frac{\alpha c}{v}\right)^2\right]. \quad (10.5)$$

This gives a shift of 28 A.

Lindholm's theory assumes that the perturber can be localized sufficiently long so that it is reasonable to consider a polarized atom. This is in contrast with our calculation, which assumes the particles to be moving so quickly that the atom is at all times essentially free. Also our treatment which uses the Born approximation is restricted to perturbers with kinetic energy large compared to the atomic energy level separations (e.g., $\Delta\nu_{4s,4p}$).

These two extreme assumptions predict similar effects, since the physical situations being considered lie between their separate provinces. Their agreement makes the magnitude of the result more secure, especially since they are reasonably close to the experimental value.

ACKNOWLEDGMENT

This paper evolved from a technical memorandum written two years ago by Margenau and Landwehr. To Dr. Landwehr we are indebted for the first draft of Secs. II and IX. Dr. Kivel's contribution (see Sec. X) is gratefully acknowledged. After this article had been written, a manuscript and a research memorandum by M. Baranger were received. Unfortunately, their contents, though of great relevance to the problems here discussed, could not be included in this survey.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

Electronic Band Structure of Solids by X-Ray Spectroscopy

L. G. PARRATT*

Cornell University, Ithaca, New York

CONTENTS

	Page
I. Introduction	616
II. General Aspects	617
Advantages and Disadvantages	617
Two Energy Regions	618
III. Resolving Power and Corrections	618
Corrections to Observed Spectra	619
IV. Energy Level Diagram	620
Terminology	620
"One-Electron-Jump" Diagram	621
X-Ray Excitation States: BEE and VEC Types	622
X-Ray Energy Level Diagram	626
Absorption Spectra: Extended Structure	628
Widths and Lifetimes of X-Ray States	630
V. Experimental Curves and Interpretations	632
Sodium, Magnesium, and Aluminum	632
Nickel, Copper, and Zinc	636
Potassium Chloride	639
Lattice Spacing and Phonons in KCl	640
Excitation States for KCl in Absorption	643
Other Types of X-Ray Curves	644
VI. Further Reading	645

I. INTRODUCTION

X-RAY spectroscopy provides the most precise and reliable information about the general electronic structure of atoms. This method is likewise powerful in elucidating certain features of the general electronic structure of solids. Metals, semiconductors, and insulators may all be studied alike.

In principle, x-ray emission spectra give a more or less direct view of the outer occupied electronic levels or, rather, of the corresponding energy states of the solid, e.g., the valence band of states; and absorption spectra provide a similar view of the states corresponding to the unoccupied levels. With a combination of emission and absorption spectra, one hopes to see the forbidden energy gap characterizing nonmetals, and to see the Fermi energy for metals. However, what one actually sees is often dominated by the local discrete states set up as a consequence of the inner electron vacancy. The most difficult feature of this x-ray method is that the unperturbed or normal solid-state band spectra and the local discrete state spectra are superimposed, and careful experimental measurements and a careful interpretation are required. The interpretation is much more difficult than has been generally recognized, but also offers greater possibilities.

Another feature of the method is that, in interpreting emission or absorption spectra, due account must be taken of the relative transition probabilities. These probabilities involve the wave-function characters (symmetries) of the initial and final states, in accord with the well-known selection rules. Although it is commonly supposed that the admixture of the various wave-function symmetries is rather great in the valence and conduction bands of solids, it is still not complete. However, with a combination of spectra characterized by two or more different high-energy states, for example with the $K(1s)$ and the $L_{III}(2p)$ spectra, we can determine (again in principle since the interpretation is very difficult) the energy positions, widths, and shapes of the energy levels or bands, and the degree of occupancy of the corresponding outer electronic levels or bands, of any amenable solid.

For the investigation of its x-ray emission spectra the solid specimen is prepared as an anode of the x-ray tube, or as a secondary radiator for fluorescence radiation. For absorption spectral measurements it is prepared as a thin x-ray absorber. All the elements in the periodic table (except hydrogen and helium) can be studied by this method, although for many of them the variety of practical high-energy states is restricted.

Wavelengths employed in this work range from about 0.5 A (~25 000 ev) to about 800 A (~15 ev). There is thus brought into play a wide variety of both old and new experimental techniques, techniques involving the radiation source, diffraction of the beam for monochromatization, dispersion and resolving power, and devices for precise measurement of the x-ray relative intensity.

It is to be understood that this brief review is not intended to be exhaustive; it is intended to introduce and to orient the fundamental aspects of the subject in its present state of development for those physicists, chemists, metallurgists, etc., who have not had time to keep up with it. Other reviews of the subject have recently appeared[1–6] emphasizing the historical de-

* This research was supported by the U. S. Air Force through the Air Force Office of Scientific Research of the Air Research and Development Command, under contract No. AF18(600)-300.

[1] A. E. Sandström, "Experimental methods of x-ray spectroscopy: ordinary wavelengths," in *Handbuch der Physik*, S. Flügge and Marburg, editors (Springer-Verlag, Berlin, 1957), Vol. 30, pp. 78–245.

[2] D. H. Tomboulian, "The experimental methods of soft x-ray spectroscopy and the valence band of the light elements," in

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

velopment and most of the experimental techniques, so we place our greatest emphasis upon current (and perhaps future) problems of interpretation.

Before launching into typical x-ray spectra and the interpretational problems, we first (a) mention the more general aspects of the method, (b) emphasize some experimental factors that must be appreciated before interpretation should be undertaken, and (c) point out certain essential features of the x-ray energy level diagram that have so far received inadequate attention.

II. GENERAL ASPECTS

Advantages and Disadvantages

As compared with other methods for studying the electronic structure of solids, the x-ray spectroscopic method has five significant general advantages:

(1) All types of solids can be studied (metals, semiconductors, insulators, pure elements, compounds, alloy systems, etc.).

(2) Details of the structure are obtained by the use of a relatively narrow high-energy state as a scanning probe.

(3) A wide energy range of the structure is covered giving widths, shapes, and energy positions of many states and bands of states.

(4) The degree of admixture or hybridization of wave-function symmetries is revealed by the use of two different types of inner electron vacancies.

(5) The spectra are relatively insensitive to the chemical impurities, surface states, or other types of lattice defects (except the defect due to the inner electron vacancy itself) which are so troublesome in most solid-state work.[7,8] Two important general disadvantages of the method are (a) the theoretical and experimental techniques for quantitative work are rather exacting, and (b) the problems of separating the two superimposed types of spectra, *viz.*, the unperturbed solid-state band spectra and the local discrete state spectra, have not been generally solved.

The theoretical aspect deserving special mention is the atrocious many-electron nature of x-ray transitions. Numerous simplifying approximations must be adopted for reasons of expediency and the final effects of these approximations remain unknown. Comment must also be made on the very short times involved in x-ray transitions,[9] of the order of 10^{-12} to 10^{-16} sec, making it important in most instances to take time-dependencies into account in the theoretical descriptions of the pertinent states and transitions. On the experimental side, the measurements must be made with the best techniques and with the highest accuracy in order to allow the necessary corrections to be applied with reasonable conviction. These corrections are listed later. Many spectra have been recorded in the literature for which the appropriate corrections are very large and have not been made.

With regard to the second disadvantage, studies of the local discrete states themselves (which we shall call "excitation" states[10]) provide valuable information about the perturbation due to the inner vacancy; such a perturbation becomes an additional tool in solid-state analysis. The excitation states are most easily recognized in absorption spectra with semiconductors and with insulators, although they are probably present also in some emission spectra. They are very much akin to those commercially-important states (in optical fluorescence) associated with an impurity atom.[11]

Handbuch der Physik, S. Flügge/Marburg, editors (Springer-Verlag, Berlin, 1957), Vol. 30, pp. 246–304.

[3] C. H. Shaw, "The x-ray spectroscopy of solids," in *Theory of Alloy Phases* (American Society of Metals, Cleveland, 1956), pp. 13–62.

[4] Y. Cauchois, *Les Spectres de Rayons X et La Structure Electronique de la Matière* (Gauthier-Villars, Paris, 1948).

[5] H. Niehrs, Ergeb. exakt. Naturw. 23, 359 (1950).

[6] Landolt-Börnstein Tables (Springer-Verlag, Berlin, 1955), sixth edition, Vol. 1, Part 4, pp. 769–867. Many experimental curves are reproduced without any critical comment.

[7] Since the inner electron vacancy (the essential vacancy for an x-ray transition) involves a relatively high energy, its production is not much affected by the low energies that characterize the bonds in the crystal lattice. It is most improbable that an appreciable fraction of the observed x-ray transitions takes place in the immediate neighborhood of a chemical impurity or other lattice defect (exclusive of the defect due to the inner vacancy itself) unless these defects are present in an *unusually* high concentration.

[8] Another type of impurity, however, is a serious problem. This is the matter of being sure that the material under study is really of the crystalline form desired and is really free of volume and surface contaminants. In emission from an x-ray tube anode, only the first hundred or thousand angstroms of material is sampled, a sampling depth that depends upon the material, geometry and voltage of the anode. The effective temperature of the emitting volume of the anode is difficult to know or to control, especially with a nonmetallic anode, and some physical or chemical change may be unwittingly induced. A thin layer of oxide or sulfide, or a time-dependent layer of tungsten or of a carbonaceous deposit on the anode surface, is often a trial. In x-ray emission in fluorescence, the sampling depth is somewhat greater, the temperature is much easier to know and to control, and generally the physical and chemical state is much better known.

[9] An initial x-ray state may decay either radiatively or nonradiatively. If it decays nonradiatively, an Auger electron, rather than an x-ray photon, is emitted. The probability for radiative decay is called the fluorescence yield (it should be called the "radiative" yield). Values and discussion of this yield are given by Broyles, Thomas, and Haynes, Phys. Rev. 89, 715 (1953); E. H. S. Burhop, *The Auger Effect* (Cambridge University Press, London, 1952). For any state of energy less than about 15 kev, the Auger process is the more probable one and, hence, dominates the state lifetime; and for states of even smaller energy, the radiative probability becomes rapidly less, being less than 0.1 for states of energy less than about 4 kev.

[10] The term "exciton" is commonly used in reference to one type of excitation states in absorption spectra, especially in ultraviolet absorption. "Exciton" is a theoretical term that refers specifically to a coupled electron and positive hole that, as originally proposed by Frenkel, travel as a unit more or less freely around in the lattice. This term has acquired so much of the aura of the theoretical one-electron model that its use is apt to be misleading in practice, especially in the interpretation of x-ray spectra. Further comments on the one-electron model are made later.

[11] The "impurity atom" in the present case is commonly spoken of as one of atomic number $Z+1$ in a lattice site normally containing an atom of atomic number Z. Actually it is somewhere

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

Excitation states are discussed further in connection with the energy level diagrams.

Two Energy Regions

Requiring high resolving power (an energy resolution of about 0.1 ev) and accurate measurements of relative intensity (of the order of one percent), this spectroscopic method divides experimentally into two energy regions, with a strong tendency for a corresponding dichotomy of the subject as it appears in the literature.

The division is essentially according to the type of monochromator: For the region roughly 0.5 to 18 A, reasonably perfect crystals (calcite, quartz, mica, etc.) are used either on a two-crystal spectrometer (sometimes called a double crystal spectrometer) or on a focusing instrument in which a single crystal is carefully bent to focus the diffracted beam. For the region roughly 15 to 800 A, ruled concave gratings are used at grazing incidence in a Rowland mount. Vacuum instruments, or instruments whose x-ray path (including any transmission windows) is otherwise made easily penetrable, are employed for wavelengths greater than about 2.8 A.

Another difference between the two regions has to do with the method of measuring the x-ray relative intensity: The precise and reliable photon-counting techniques, especially with the proportional counter,[12] are used most advantageously in the 0.5 to 18 A range. The less accurate photographic techniques are used in the longer wavelength range. However, much work in the former range, especially pioneering work, has been photographic, and some success has been achieved with photon counters in emission work in the latter range. In either range, the photographic method is far better suited for quick surveys of wide spectral regions since only a single exposure is required. It is well known, of course, that the photographic method is well adapted to precise determinations of wavelengths of *well-resolved symmetrical* lines in either range, but a difficult and often impractical procedure must be set up for each exposure in order to translate the photographic blackening to a reliable intensity (or absorption) scale. This fact seriously hampers accurate photographic studies of such spectral properties as widths, shapes and relative intensities. Indeed, reports of such properties, if a reliable scale has not been established, are of very little or no quantitative value.

A third difference between the two regions is in the type of radiation used in absorption work. Continuous x-radiation from a conventional type x-ray tube is evidently too feeble to be practical at the longer wavelengths. Work to date, with one exception, has been done with optical spark lines. An excellent source of continuous radiation is now available from high-energy centripetally accelerated electrons in the orbit in a synchrotron.[13,14] This type of orbit radiation may soon replace the radiation from conventional x-ray tubes in x-ray spectroscopy in *both* wavelength regions, the beam being used directly in absorption work and being used to produce by fluorescence the characteristic emission spectra of any material placed in the beam.

Incidentally, the availability of the continuous radiation from the synchrotron "x-ray tube" will no doubt further stimulate the development of photon or electrical counters for the long wavelength region since this type of measurement system now becomes useful in absorption work as well as in emission. Because of the short pulses of radiation from the synchrotron, such counters will probably have to be arranged to measure an average number of photons per second by means of some integrating device.

III. RESOLVING POWER AND CORRECTIONS

In either energy region, two kinds of resolving power are involved: first, that due to the so-called "spectral window" of the instrument,[15] and, second, that due to the effective width and shape of the high-energy state in the x-ray transition. The central region of the spectral window, in the case of the two-crystal spectrometer, is determined essentially by the lattice type and degree of perfection of the monochromator crystals.[16] The tail regions of the window are usually

between Z and $Z+1$, depending upon the particular inner vacancy, because the screening of an inner electron on an outer electron is less than unity.

[12] The proportional counter, if designed to have a very stable and narrow pulse-height distribution, may be used with a single-channel discriminator to reduce the background intensity and also the intensity of any higher order spectra.

[13] D. H. Tomboulian and P. L. Hartman, Phys. Rev. **102**, 1423 (1956); D. H. Tomboulian and D. Bedo, J. Appl. Phys. **29**, 804 (1958).

[14] L. G. Parratt, Rev. Sci. Instr. **30**, 297(L) (1959).

[15] "Spectral window" refers to the effective *over-all* spectral response of the instrument. If the beam incident upon the spectrometer has a flat spectrum, the spectral window may be interpreted as the function which gives the probability that a photon of energy $h\nu$ will be counted when the spectrometer is set at energy $h\nu_s$. For example, the spectral window of a two-crystal spectrometer includes the effects of (a) the geometry of the horizontal and vertical collimating slits, (b) fluorescent radiation and the diffuse or Compton or other spurious scattering (especially from the second crystal), (c) any higher order spectra, (d) any frequency-dependent attenuation of intensity in the x-ray path in the spectrometer, (e) the relative wavelength sensitivity of the intensity-measurement system, as well as (f) the diffraction pattern of the two crystals in the $(n_A, \pm n_B)$ position of operation. The true spectrum is distorted by the spectral window by an amount that depends upon the window's width *and shape*—the somewhat remote tails of the window's shape cannot be ignored.

The dispersion of the instrument, *viz.*, the spread of an x unit or of an ev in Bragg angular measure or in linear measure along a photographic plate, should not be confused with resolving power, although with some types of instruments dispersion and resolving power are closely related.

[16] Calcite crystals are the traditional favorites, but recently quartz crystals have been shown to have higher resolving power and comparable percent reflection. [Adell, Brogren, and Haeggblom, Arkiv Fysik **7**, 197 (1954)]. Practical attainment of the higher resolving power requires skillful operation of a well-designed instrument. For example, with the two-crystal spectrometer with good quartz crystals, angular increments as small as 0.1 sec of arc must be measured accurately and the axes of the two crystals must be maintained parallel to within a few seconds of arc for *all* angular settings including both the parallel and the antiparallel positions of the instrument.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

determined by a combination of the so-called thermal diffuse scattering and the slit geometry. In the grating spectrograph, the window is usually dominated by the width of the slit. In the case of the focusing bent-crystal instrument, the window is due to some combination of the crystal, slit, scattering, and focusing imperfections.

In the 0.5 to 18 A range, the over-all effective resolving power is usually determined, not by the spectral window, but by the width of the high-energy state. This state width in the 0.5 to 18 A range is of the order of 0.2 to 8 ev at half-maximum, depending upon the particular state and atomic number. On the other hand, in the 40 to 800 A range, the slit is usually the more important; and it is of the order of 0.01 to 0.5 ev wide. In the longer wavelength range, the "smearing" due to the over-all resolving power is often negligibly small, but this is never so in the 0.5 to 18 A range.

The rather large width of the high-energy state in the shorter wavelength range has been considered by many investigators to be a serious disadvantage, often over-compensating for (a) the ease and efficiency of producing the x-rays, (b) the convenience of the more penetrating x-rays, and (c) the precise measuring techniques of the photon counters. Recently, however, methods have been developed[17] for correcting the observed spectra for both kinds of resolving power if (a) the experimental curves are very accurately recorded, and (b) if the width and shape of both the spectral window and the high-energy state are accurately known. With the two-crystal spectrometer the window shape (exclusive of the remote tails) is usually intermediate between Gaussian and Lorentzian; with the ruled grating it is usually more nearly Gaussian. The shape of the high-energy state is commonly taken to be Lorentzian, as due to lifetime, but it also contains a Gaussian component due to lattice vibrations and often a width due to unresolved multiplet states as well.

Corrections to Observed Spectra

In general, several different corrections must be made to the directly observed spectra in order to obtain a reasonably reliable view of the energy level structure of the solid. The most important corrections in the 0.5 to 18 A range are commonly those for resolving power, both for the spectral window and for the high-energy state. In the 15 to 800 A range, often the conversion from photographic blackening to x-ray intensity is the most important correction. In either range, other corrections are also involved. Briefly, the effects for which corrections must be made, and in the order of procedure for making them, are (a) nonlinearity of the intensity scale, (b) instrumental resolving power (spectral window), (c) self-absorption by the material emitting the x-rays, both of the emission line or lines and of the ubiquitous continuous radiation, (d) background under the particular emission contour being studied, (e) satellites (or other "spurious" lines) in emission, and (f) width and shape of the high-energy state either in emission or in absorption. It is presumed that the material emitting or absorbing the x-rays is clean and of known physical and chemical state. In case of mixed materials, correction for the foreign material must also be included in (c), (d), and (e).

The relative importance of each of these six corrections, as well as the details of the correction procedures, depends upon the particular spectrometer or spectrograph used and upon the particular spectra under study. Unfortunately, even a reasonably adequate discussion of these corrections has not been attempted in most of the published work in this subject to date, and this fact has greatly impeded progress in interpretations.

It is significant that now it is possible, in either wavelength range, albeit difficult in some instances, to obtain "experimental" curves in which essentially the product $N_a(E)T(E)$ is plotted *vs* the energy E. In this product, $N_a(E)$ is the density of those states, including the excitation states, having the appropriate wave-function symmetry for the particular transition according to the radiative selection rules, and $T(E)$ is the transition probability. [Logically, $T(E)$ includes the effects of wave-function symmetry, but it is customary, because of the use of the energy level diagram, to place the subscript s or p, etc., on $N(E)$.] If one desires a view of the normal energy band structure of the solid, $N(E)$, unperturbed by the inner electron vacancy and regardless of the wave-function symmetries of the states, corrections for two additional effects must be made: (g) excitation spectra and (h) variations in the transition probabilities, with due regard for wave-function symmetries,[18] for the pertinent transitions

[17] J. O. Porteus and L. G. Parratt (to be published). Preliminary reports appear in Bull. Am. Phys. Soc., Ser. II, 2, 55 (1957); *ibid.* 3, 263 (1958).

[18] For a free atom, the transition probability is proportional to $\nu|M_{i,f}|^2$ in emission and to $1/\nu|M_{i,f}|^2$ in absorption, where $M_{i,f}=\int\Psi_f^*(\partial/\partial x)\Psi_i d\tau$, ν is the radiation frequency, Ψ_i and Ψ_f are the wave functions of the initial and final states, and the integration is over all space. For a solid, a similar expression applies but now the wave functions, especially for a state which contains an outer vacancy (e.g., the final state in emission) or an extra outer electron (e.g., the final state in absorption), are not at all well known, and, in particular, they cannot be assumed to have one-electron properties or even a convenient symmetry for purposes of integration.

As a part of the wave function of the final state in an absorption transition, the wave function of the ejected electron, if it is ejected completely free of the positive hole it leaves behind, is commonly assumed to be that of a plane wave in the lattice. If this assumption is valid, then in the transition probability, the linear momentum of this plane wave electron must be expressed as an angular momentum about the atom from which the electron is ejected.

In emission, the transition probability is proportional to ν, and, hence, the emitted intensity is proportional to ν^2. Some authors have partially corrected for the transition probabilities by dividing the observed intensity by ν^2 (the division by ν^3 by a few authors is evidently in error).[2] By the same argument, the

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

from (or to) each and all states in the bands. With these eight corrections (a) through (h), one would obtain the density of states curve $N(E)$ *vs* E. But we do not yet know how to make corrections (g) and (h) satisfactorily. For example, each solid-state energy level or band contains an admixture of several different symmetries, and correction (h) presumes knowledge of this admixture. At present the most reliable method of obtaining this knowledge is from experimental x-ray spectroscopy with two or more high-energy states of different symmetries. However, in many respects, the degree of admixture is as interesting as the fully corrected $N(E)$ *vs* E curve itself.

Before discussing a few typical experimental spectral curves, let us prepare generally for some of the problems of interpretation. We devote considerable time to the qualitative details of the energy level diagram since it is here that confusion is most apt to arise.

IV. ENERGY LEVEL DIAGRAM

Terminology

At least to the uninitiated there seems always to be confusion among the terms "state," "energy level," "energy level diagram," "energy band," "electronic shell," "orbital," "ordinary x-ray state," "x-ray satellite state," "ionization state," "excitation state," and "electronic level." Let us first try to clarify these terms. This is a major step in the understanding of an energy level diagram.

"State" refers to an assemblage of particles in a particular configuration in a system. The system may be a gaseous atom, a polyatomic molecule, or a localized part of a large piece of liquid or solid material. In a liquid or solid, the degree of localization is determined by the lifetime of the state in question and by the Franck-Condon principle. The energy of a state is also called the energy level in implicit or explicit reference to an energy level diagram. Such a diagram is discussed in detail later. An energy band in a solid (or in a liquid or in a polyatomic molecule) refers to a band of energy levels extremely closely spaced.

Those electrons that are described by the same wave-function characteristics (except for the direction of electron spin) are said to make up, or to be in, one "electronic shell" or "orbital." For example, the $2p_{\frac{1}{2}}$ electrons (called L_{II} or sometimes L_2 electrons) comprise one electronic shell or orbital (called the L_{II} or L_2); so do the $3d_{\frac{1}{2}}$ electrons (called M_{IV} or M_4); etc. The term shell or shells is used especially for the appropriate group or groups of electrons or orbitals in a monatomic gaseous atom and for the inner electrons in a polyatomic molecule, a liquid or a solid, but not very often in reference to the outer electrons in a nonmonatomic system.

It follows that each state of a many-electron assemblage—a gas, liquid, or solid—refers to a configuration of *many* constituent electronic orbitals. It also follows that only an electronic shell or orbital (but *not* a state or band of states or an energy level) can be said to be occupied or unoccupied, i.e., by one or more electrons.

An ordinary x-ray state always refers to an electron configuration in the system in which there is one and only one inner electron vacancy. For this reckoning, any bound electron in the system, except a valence electron,[19] is an inner electron. This definition of an ordinary x-ray state unfortunately excludes the state corresponding to a vacancy in the valence orbitals. In spite of this exclusion, the valence state is, for obvious reasons, always included in the x-ray energy level diagram of ordinary levels.

Two or more inner vacancies correspond to a second- or higher x-ray state, sometimes called an x-ray satellite state.[20] Higher order or satellite states are usually emphasized only in the emission of satellite lines, but they are also of direct interest in absorption spectra.

Any ordinary x-ray state may be an ionization state or it may be one of various types of excitation states. It is an ionization state if the ejected electron is completely free from the Coulomb field of the atom (or ion) in question, and if there is no abnormal configuration of the valence electrons. If the inner electron is ejected from its shell but is not completely free, it remains bound in an orbital outside of the occupied valence orbitals. In this event the system is in one type of excitation state. Excitation states are discussed in detail later.

In this terminology, we refer to the type of many-electron assemblages actually encountered in the interpretation of any and all x-ray spectra. Insofar as the gross features of the *inner* part of the atom are concerned, an x-ray state may be thought of as a "one-electron vacancy state." One of the theorems in atomic spectroscopy says that a one-vacancy system behaves in many respects like a hydrogenic or one-electron system. In hydrogen, it suffices to treat the energy of the system as being the same as the energy of the one electron, and no distinction is made between the electron energy and the system energy. Because of the difficulty of theoretical calculations of a many-electron

observed absorption spectra (properly plotted as absorption coefficients) should be multiplied by ν. These partial corrections are of little importance in the short wavelength range but often change considerably the general shape of either emission or absorption curves in the long wavelength range.

[19] We use the term "valence electron" to refer to any electron in the *outermost* band of electronic orbitals that is normally partially or fully occupied. In this use, we do not separately distinguish a conduction electron in a metal, nor heed the special role of certain valence electrons in the physical or chemical bond in molecules or solids.

[20] This is the conventional definition of an x-ray satellite state, but the exclusion of a valence vacancy as the second vacancy does not mean that the energy of the state is not affected thereby. In fact, if the second vacancy is in what we later call an excitation valence orbital, especially if it has s-type symmetry, it produces a very significant change in the energy of an inner-vacancy state. We shall later refer to such an altered state as one type of x-ray excitation state.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

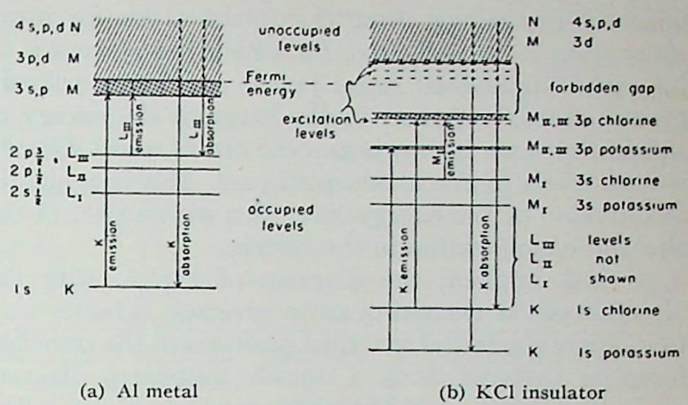
system, the theory of the one-electron approximation is used as the basis of the theory for the many-electron atoms and for the solid state. It is common in solid-state lore to refer to a change in a single electronic level as being simply the change in the energy of the system. This, of course, would be valid if, as *the* electron shifts in its momentum or energy position, there is no shift of any other electron in the system. Actually, in most physical phenomena when examined in some detail, including x-ray spectra, the interaction between electrons is so great as to invalidate this one-electron approximation.

The symbol terminology of the one-electron model, *viz.*, *s*, *p*, *d*, etc. (lower-case letters) to indicate electrons, electronic orbitals, bands of orbitals, states, energy levels, and energy bands, is usually retained in x-ray spectroscopy in spite of the intrinsic many-electron nature of the problems.[21] These one-electron symbols are retained with the implicit understandings that (1) they refer to the *principal* jumping electron in the x-ray transition, (2) they do not indicate an absence of interaction between this principal electron and other electrons in the system, and (3) in applying the radiative selection rules to determine the allowed transitions, the assumption or approximation is made that the change in the angular momentum of the entire system is given by the angular momentum change of just the principal jumping electron, i.e., that no other electron changes its angular momentum. It must be realized that retention of the one-electron symbols carries along a tendency to confuse the many-electron system energy levels with the one-electron system energy levels; *every effort must be made to avoid this confusion*.

"One-Electron-Jump" Diagram

Figure 1 presents a rough sketch of a conventional so-called energy level diagram (a) for a metal and (b) for an insulator. Although this type of diagram contains some confusion and is inherently inadequate, as implied generally above and as pointed out specifically below, it is the one that has been most used in this subject to date and therefore deserves discussion.

By definition, each level in the diagram is supposed to denote the energy of the entire system when one electron is missing from the particular electronic shell or orbital specified in the energy level terminology. It is supposedly a diagram of x-ray ionization states. A few so-called excitation levels are sketched in for the case of the insulator although, as we shall see, inclusion of these states in this type of diagram is somewhat abortive. The principal feature intended, of course, is that the transition arrow give the energy of the photon



(a) Al metal (b) KCl insulator

FIG. 1. Conventional x-ray "energy level" diagram (arbitrary scale) for each of two solids. This type of diagram was devised for a combination of a one-electron-vacancy model and a one-electron model; it is incomplete and misleading (see text).

emitted or absorbed. With a linear energy scale, this energy is interpreted as being the length of the arrow; however, for convenience in x-ray spectroscopy where a wide range of energies is involved, a logarithmic scale is usually used.

The confusion in this diagram appears in both atomic and solid-state work, but it is in solid-state work that it becomes especially serious.[22] We shall presently propose another type of diagram, one that is more complex but that, we believe, has the necessary features to avoid the pitfalls of misinterpretation inherent in the diagram of Fig. 1.

One type of confusion is the following. The lower part of the "occupied" region of the diagram is just an inversion[23] of the diagram conventionally used as a satisfactory approximation in interpreting the gross features of x-ray spectroscopy of the "inner" or high-energy levels. But, in the "occupied" and "unoccupied" regions near the valence and conduction bands, we have deviated from the one-vacancy system and have drawn in the normal or unperturbed solid-state bands in which the effect of the vacancy is ignored. In fact, we usually treat the upper or solid-state part of the diagram as a one-electron system instead of as a one-vacancy system. This composite leads to difficulties when transitions between the lower and upper parts are considered. One difficulty becomes readily apparent

[21] Upper case letters are often used in referring to theoretical x-ray satellite states corresponding to multiple inner vacancies. In this case the electron configuration is presumed to be completely known and to behave as a monatomic many-electron system.

[22] It may be argued that, in a general sense, any diagram is merely a shorthand aid to memory, and if someone finds a particular scheme useful as a memory aid, it cannot be properly called inadequate or inconsistent for him. But diagrams somehow have a penchant for becoming pedagogical and interpretational aids as well as memory aids, and a poor diagram is often definitely harmful.

[23] All the bound stationary states of a hydrogenic or one-electron system are of negative energy, and all unbound or continuum states are positive. In the case of a one-vacancy system, the bound states are properly reckoned as positive. Accordingly, several years ago, before "solid-state spectroscopy," the x-ray diagram was drawn with the *K* state at the top instead of at the bottom. The arrangement with the *K* state at the top is conceptually better for "atomic" x-rays but it does not allow conveniently for the continuum of states. The inverted arrangement of Fig. 1 has been used in deference to the convention in solid state work.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

when it is realized that the initial state for x-ray absorption is actually one for which the system is in its normal unexcited state, i.e., in its "single-valued" ground state,[24] whereas in the diagram the energy of this initial state is drawn as some one of *many different* energy levels in the unoccupied part. This is evidently a confusion of the energy level idea with a part of the electron configuration in the system.

As just implied, the diagram of Fig. 1, with the direction of the transition arrow reversed, is better used to indicate the initial and final positions of the principal jumping electron. Such a one-electron-jump diagram *could* purport to give the energy values of the initial and final states on the argument that each electron configuration, when completely specified, corresponds to a definite energy of the system. The difficulty in this use of the diagram is that there are generally many different electron configurations associated with each inner vacancy, and there may be several different electrons jumping in each transition. For example, it is clear that an excitation energy level exists only because of, and during the lifetime of, the inner electron vacancy. Consequently, when an electron jumps from any occupied orbital to fill the inner vacancy, the excitation level ceases to exist (or changes to a different excitation level), and any electron in the excitation orbital must also jump to some new orbital. In the special event that the inner vacancy is filled by the excitation electron, i.e., by the electron in the excitation orbital, the final vacancy is in no sense in the excitation energy level as this level appears in the diagram. Energy-wise, where did the electron jump from? No transition arrow can be drawn in the conventional diagram to indicate this jump.

X-Ray Excitation States: BEE and VEC Types

A proper x-ray energy level diagram should have a single energy level corresponding to the ground state (which is conveniently called zero energy) as the initial state for each and every x-ray absorption transition; and the final state, being one in which the ejected electron is in any one of many otherwise unoccupied outer electronic orbitals or in the continuum, should be represented as one of *many* high-energy levels. For example, for K absorption, the diagram should show the variety of final-state energies as a type of splitting of the conventional K state, including the continuum of K levels as well as the several discrete K levels. Any one of these many K levels may be the terminus of the head of the arrow representing a K-absorption transition. Similarly, each one of the other conventional x-ray energy levels (L_I, L_{II}, L_{III}, M_I, M_{II}, etc.) should be split into several discrete levels and a continuum. An x-ray state in which the ejected electron resides in one of the bound orbitals is one type of excitation state. To the many possible states of this sort, we assign the descriptive name "bound-ejected-electron" (BEE) excitation states.

Bound-ejected-electron excitation states are commonly produced in x-ray and ultraviolet absorption but are believed to be only rarely produced by bombardment with cathode electrons in an x-ray tube anode. Such states probably dominate the irregular structure observed near the absorption edges, e.g., see Fig. 2 for a curve for a monatomic gas,[25] Fig. 3 for a polyatomic gas, and Fig. 4 for a solid. But, especially in solids, other types of excitation states may also be involved. Since 1920, any observed structure near the "main" absorption edge has been known as Kossel structure.[1,26]

Another type of excitation state was implied earlier, *viz.*, a state in which the system has some abnormal

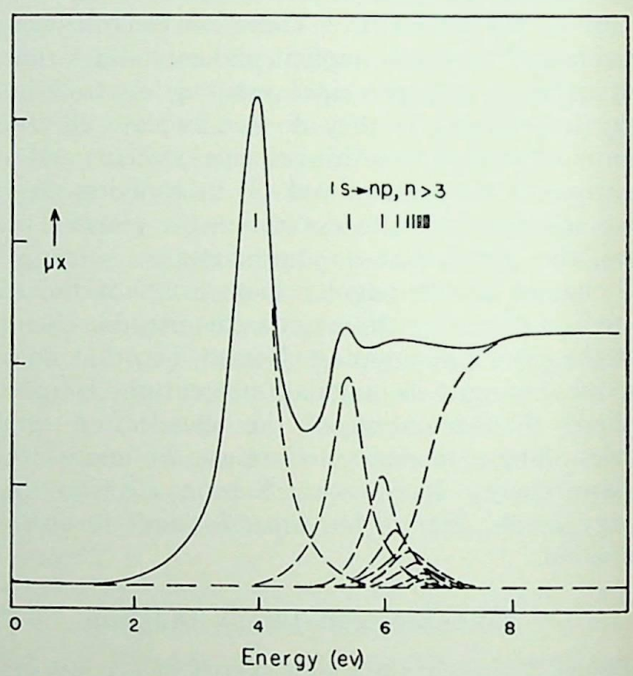


FIG. 2. K-absorption spectra (uncorrected) for gaseous argon [L. G. Parratt, Phys. Rev. **56**, 295 (1939)]. The structure is dominated by bound-ejected-electron excitation states; the "main" absorption edge occurs at the series limit of these states.

configuration of the valence electrons. For many materials, it is highly probable that only the so-called normal configuration of valence electrons is significantly involved. But for some materials, e.g., solids having s-type valence electrons and perhaps an incomplete inner shell (transition metals), the "abnormal" con-

[24] The energy of this ground state is conveniently taken as the zero energy for all kinds of systems—gases, liquids, or solids.

[25] The correction for instrumental resolving power as reported by L. G. Parratt [Phys. Rev. **56**, 295 (1939)] is not quite right as judged by the results of the new correction procedure. The curve in Fig. 2 is not corrected for resolving power.

J. A. Soules and C. H. Shaw [Phys. Rev. **43**, 470 (1959)] also report this absorption spectrum but they evidently have an inadvertent "thickness effect" in their curve. Indeed, many or most of the x-ray absorption curves reported in the literature suffer from this effect. For a discussion of the thickness effect, see Parratt, Hempstead, and Jossem [Phys. Rev. **105**, 1228 (1957)].

[26] A. H. Compton and S. K. Allison, *X-Rays in Theory and Experiment* (D. Van Nostrand Company, Inc., New York, 1935), pp. 662–671.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

figurations are believed to be important. Let us look at the valence electrons more closely.

As the inner vacancy is being produced, a change occurs in the Coulomb field in which all the electrons of the system find themselves. No doubt, the final configuration of all the electrons would be essentially unique (i.e., only the normal configuration) for a given inner vacancy *if* the inner vacancy were formed very slowly and if the electrons were granted a long time in which to readjust to the altered field. But such deliberation is not allowed. The change in the Hamiltonian of the system occurs in less than about 10^{-18} sec, and in consequence of this change all the electrons are forced to change their wave functions. They do so as rapidly

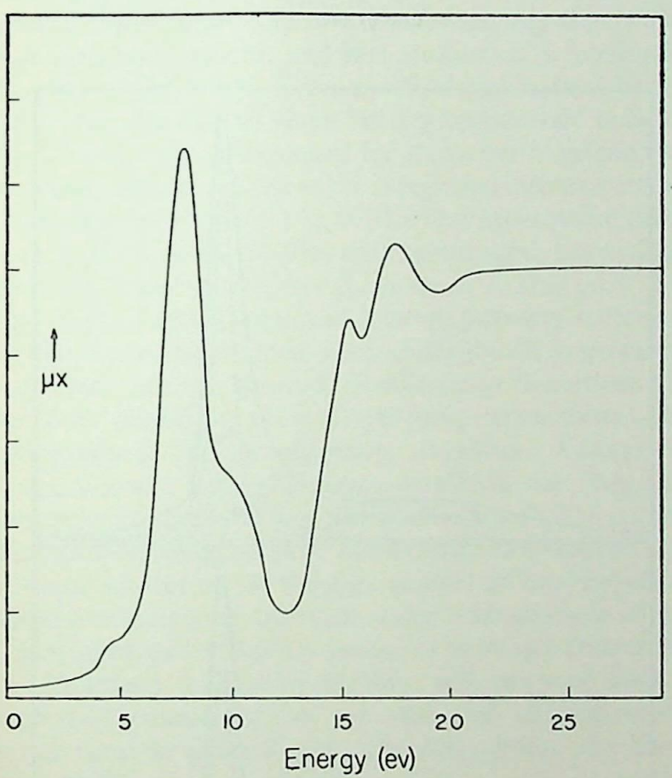


FIG. 3. *K*-absorption spectra (uncorrected) for gaseous chlorine [Stephenson, Krogstad, and Nelson, Phys. Rev. 84, 806 (1951)].

as they are able. Generally, it may be presumed that the inner electrons shift rather quickly to the respective new orbitals closer to the nucleus of the atom in question, but the slow-moving valence electrons may require more time. The alloted time is about the lifetime of the inner vacancy, of the order of 10^{-16} sec for K states. At the termination of this allotted time some of the valence electrons may have succeeded in following the new orbitals and the others may have been obliged to go to an unoccupied excitation orbital[27] or to the unoccupied part of the continuum. The energy of the system depends upon the particular configuration

[27] To illustrate, the 4*s* valence electron of monatomic gaseous potassium may go, upon production of the 1*s* hole, to the 5*s* orbital of the *new* system (i.e., according to the new Hamiltonian) leaving a vacancy in the new 4*s* orbital.

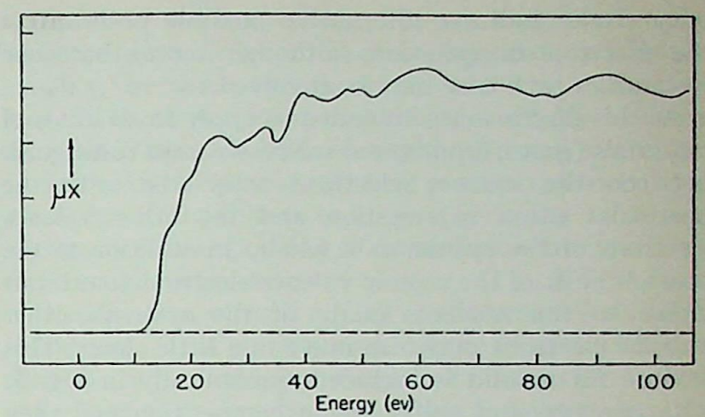


FIG. 4. L_{III} absorption spectra (uncorrected) for metallic silver [L. G. Parratt, Phys. Rev. 54, 99 (1938)].

of these electrons, and different configurations yield different energies. Without commenting further about their relative production probabilities,[28] we shall call these possible different states "valence-electron-configuration" (VEC) excitation states. They are reckoned as excitation states whether or not the inner electron is ejected completely free of the atomic field, although of course the energy depends also upon the particular final position of the ejected electron.

Figure 5 is a rough sketch which symbolizes in simple fashion the type of different electron configurations involved in the bound-ejected-electron and valence-electron-configuration excitation states for a solid. The

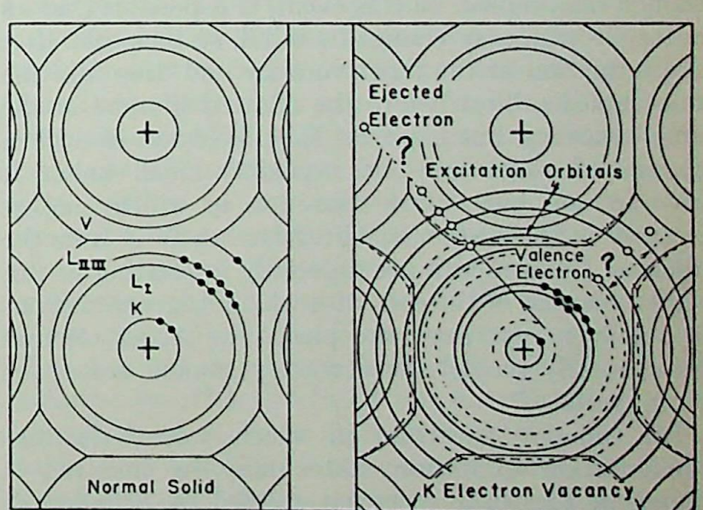


FIG. 5. Symbolic representation of different electron configurations in x-ray excitation states. For clarity, a restricted number of electron orbitals is shown: More than one valence electron and more than one ejected electron may be involved.

[28] There are many eigenvalues in the solution of Schrödinger's equation for a many-electron system following an abrupt change in the Hamiltonian, and we have mentioned only a few of them. This is done with the implied assumption that only a few have significant probabilities of production.

Also, we have chosen to use a semiclassical description of the states (in terms of specific electrons), rather than a general quantum mechanical description; the latter, although formulated exactly in equations, does not have much practical meaning without solution of the as-yet-unsolved many-body problem. Partial solutions of the quantum mechanical problem, using the best-known wave functions, are nevertheless very much needed.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

open circles indicate alternative possible positions of the electron in question, although more than one electron in each case may be involved.

As this discussion is intended to apply to all kinds of materials (gases, liquids, and solids) we must distinguish between the valence electrons associated with the particular atom in question and the other valence electrons of the system as a whole. In addition to the possible shift of the atomic valence electrons to orbitals closer to the nucleus, many of the system's other valence electrons may also move in a little closer. This feature for a solid is included symbolically in Fig. 5. The two types of shifts are in competition and they interact with each other since the objective of each is to shield the charge of the positive hole and thus to destroy the shift motivation. For example, insofar as the far-out excitation orbitals are concerned (corresponding to discrete energy levels in, or near the bottom of, the continuum), the in-flow of system electrons in a metal is expected to be much more important than in an insulator. The shielding thus effected by the system's electrons in a good metal is expected to inhibit especially the relative importance of the bound-ejected-electron excitation states in absorption,[29–31] as well as to alter the relative importance of the valence-electron-configuration excitation states in both absorption and emission.

But whether the system's valence electrons are thus effective or not, the new valence orbitals may be present though unoccupied. In this event, it is possible that an outer electron may change its mind, so to speak, after the formation of the inner vacancy and drop into the unoccupied orbital before the allotted lifetime of the inner vacancy has expired. The likelihood of such a change of mind may be negligibly small unless it involves an Auger-type transition in which another electron in the system is shifted farther away from the nucleus. Good metals are especially interesting in this regard because of the ease with which Auger transitions of small energy may take place, the Auger electron being merely moved the necessary amount across the Fermi surface.[32]

An emission spectrum in which valence-electron-configuration excitation states may be involved is shown in Fig. 6, a spectrum studied by Schnopper[33] with a two-crystal spectrometer. This spectrum is the $K\beta_{1,3}$ region for metallic manganese, the transitions supposedly being $K \rightarrow M_{II,III}$ where the $M_{II,III}$ electrons occupy the shells just inside the partially complete $M_{IV,V}$, N_I shells and the valence electrons in this metal. The directly observed spectral contour (dashed curve) has been corrected for the instrumental

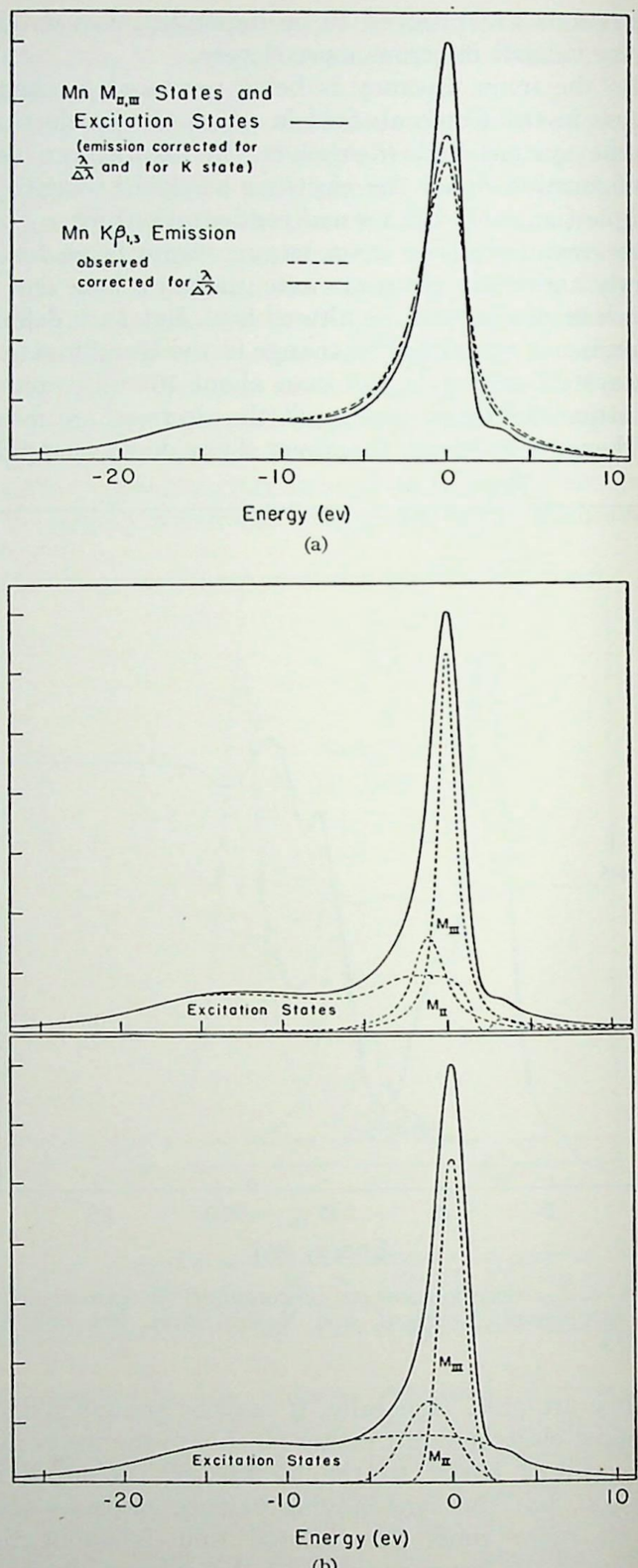


FIG. 6(a). The observed $K\beta_{1,3}$ emission and the $M_{II,III}$ states for metallic manganese. The observed contour is corrected first for the instrumental resolving power and then for the K state to give the M density-of-states contour.

FIG. 6(b). The M density-of-states contour of Fig. 6(a) resolved arbitrarily into conventional M_{II} and M_{III} components, the residue being indicated as excitation states. In the upper figure the M_{II} and M_{III} components are Lorentzian in shape; in the lower figure they are Gaussian.

[29] Y. Cauchois and N. F. Mott, Phil. Mag. **40**, 1260 (1949); and R. Landshoff, Phys. Rev. **55**, 631 (1939).
[30] J. Friedel, Phil. Mag. **43**, 153, 1115 (1952), Proc. Phys. Soc. (London) **B45**, 769 (1952); Advances in Phys. **3**, 446 (1954); J. C. Slater, Phys. Rev. **98**, 1039 (1955).
[31] L. G. Parratt and E. L. Jossem, Phys. Rev. **84**, 362 (1951), Phys. Rev. **97**, 916 (1955); J. Phys. Chem. Solids **2**, 67 (1957).
[32] Small-energy Auger transitions in a metal are discussed by P. T. Landsberg [Proc. Phys. Soc. (London) **A62**, 806 (1949)].
[33] H. W. Schnopper, M.S. Thesis, Cornell University (1958).

CC-0. Mumukshu Bhawan Varanasi Collection. Digitized by eGangotri

resolving power,[34] and then corrected for the "resolving power" due to the K state[35] by the method of Porteus and Parratt.[17] The solid curve thus represents the product of the density of p-type states and the transition probability, *viz.*, $N_p(E)T(E)$. The $N_p(E)T(E)$ curve is then shown in the lower two parts of the figure as resolved into three components, M_{II} and M_{III}, and the remaining broad low-energy component indicated as excitation states. This resolution has been carried out first with the assumption that the $M_{II,III}$ states each has a Lorentzian shape, and, then, a Gaussian shape.[36] The excitation states are commonly referred to as the final state for the β_1 line.

The wavelength of a so-called $K\beta'$ component has been reported from photographic measurements by several investigators as near the low-energy side of the low broad component, and this component is commonly known as the β' line. Three possible explanations of the origin of this β' line have been proposed but none of them seems able to account for its rather large relative intensity, about 1.4 times the integrated intensity under the β_1 line as shown in Fig. 6. The first explanation takes it to be a complex satellite corresponding to the double-electron-jump transitions $KL_{II,III} \rightarrow L_IM_{IV,V}$.[37,38] No calculations of the expected relative intensity according to these transitions have been made, but it is generally expected that, in general, double-jump transitions are far less probable than single-jump transitions. For manganese, the single-jump satellites $KL_{II,III} \rightarrow L_{II,III}L_{II,III}$, *viz.*, the $K\alpha_{3,4}$ satellites, are only one percent as intense as the parent $K\alpha_1$ line.[39] The second explanation assigns the β' component to Compton and Raman scattering of the $\beta_{1,3}$ doublet as this radiation attempts to emerge from the x-ray tube anode in which it is generated.[40,41] This requires far more self-absorption in the anode than is reasonable, and, anyway, such a scattered component is not observed accompanying x-ray lines in general, e.g., the $K\alpha_{1,2}$ lines. The third explanation is that splitting of an x-ray state may occur as a result of interaction (angular momentum coupling) between the inner vacancy and an incomplete outer shell of electrons, e.g., the M_{IV} or M_V shell for the transition elements scandium through nickel.[39,42] In general, this type of splitting is expected to be very small and to contribute at most only to the width and asymmetry of the level or of an emission line involving the corresponding state.

It seems that most if not all of the broad low-energy component in Fig. 6 arises from valence-electron-configuration excitation states. Preliminary calculations have shown[43] that, according to this interpretation, the energy intervals and the relative intensity may be of the correct order of magnitude. It is suggested that the observed relative smoothness of this component, i.e., the absence of much obvious structure, is due to the profusion of very closely spaced excitation states and to the continuum of energies of Auger electrons in a metal.

A similar low-energy satellite, $K\beta''$, is observed accompanying the $K\beta_{2,5}$ lines for elements near manganese.[38,44,45] Both the $K\beta'$ and the $K\beta''$ satellites seem to depend somewhat alike on the chemical bonds of the emitting material. It is quite possible that most or all of this $K\beta''$ component should also be assigned to the role of excitation states of the valence-electron-configuration type. Later, it is proposed that most if not all of the asymmetry of the $K\alpha_{1,2}$ lines for these materials is also due to this type of excitation state.

It may be pointed out in our general discussion of excitation states that, in principle at least, another type of excitation state may arise due to the capture of a stray electron that may be just passing through the region, or of an electron from a neighboring atom. Stray capture is very improbable, but it might be considered under certain circumstances, especially for a metal. Capture of another atom's electron, called a cross-transition, has been discussed by several investigators, especially for insulators, but its relative probability has not been well evaluated. Finally, further comment should be made on the state in which there are two or more inner electron vacancies. As stated earlier, such a multiple-inner-vacancy state is the initial state for satellite emission and is of interest in absorption spectra. According to our terminology, these states are not excitation states unless one or more of the

[34] This correction presumes that the central part of the instrumental spectral window is given by the experimental $(1, -1)$ curve which has a full width at half-maximum of 0.6 ev and which has a shape between a Lorentzian and a Gaussian, and that the remote tails (the "flat" component) of the window are a little higher than the Lorentzian. The flat component is best obtained from an analysis of the thickness effect in an absorption curve.

[35] The K state is assumed to be of Lorentzian shape having a full width at half-maximum of 1.05 ev.

[36] In each case, the resolution was made on the assumptions (a) that the full widths at half-maximum of the M_{II} and M_{III} states are 1.2 and 2.4 ev, respectively, and (b) that the relative "intensity," or state-density, at the maximum ordinate is 4 to 1. The two M widths are estimated from the radiative and Auger transition probabilities, and from a sort of "best fit" criterion in effecting the actual resolution. The relative density follows from the ratio of quantum weights, *viz.*, 2 to 1, and from the ratio of widths.

[37] According to this transition, which Sawada[38] interprets as $(K \rightarrow L_I) + (L_{II,III} \rightarrow M_{IV,V})$, this satellite has no obvious "parent" line. Nevertheless, it is customarily called a *low-energy* satellite of $K\beta_{1,3}$.

[38] M. Sawada, Sci. Mem. Kyoto Imper. Univ. **A15**, 43 (1932).

[39] L. G. Parratt, Phys. Rev. **50**, 1 (1936).

[40] H. P. Hanson and J. Herrera, Phys. Rev. **105**, 1483 (1957).

[41] F. Bloch, Phys. Rev. **48**, 187 (1935).

[42] G. Coster and J. M. Dryvesteyn, Z. f. Physik **40**, 765 (1927); G. Ortner, Wien. Sitzungsber. IIa, **136**, 369 (1927).

[43] M. H. Skolnick, M.S. thesis, Cornell University (1959).

[44] Vainshtein, Staryi, and Bril', Bull. Acad. Sci. USSR, Phys. Ser. No. 7, **20**, 1956; E. E. Vainshtein and Iu. N. Vasil'ev, Soviet Phys. "Doklady" **2**, 207 and 251 (1957); Vainshtein, Bril', and Staryi, Soviet Physics "Doklady" **2**, 518 (1957).

[45] Unfortunately, studies to date of the shape and relative intensity of the $K\beta''$ and $K\beta_{2,5}$ lines are not reliably quantitative because the corrections as listed earlier have not been made. Careful corrections are imperative for these faint lines that straddle the absorption edge of the emitting material and that appear on a relatively intense sloping background that has absorption-edge structure in it.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

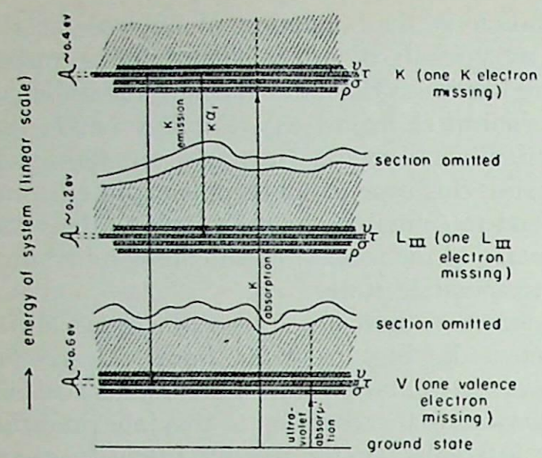


FIG. 7. Part of an x-ray energy level diagram (qualitative) for an insulator. Only a few (of many) discrete states are shown for each type of inner vacancy; each K state, for example, corresponds to a particular electron configuration of *all* the electrons in the system. All the many discrete levels in a given group, e.g., the K group, are involved in absorption spectra, but only one or a few are ordinarily involved significantly in emission spectra (see text). A different continuum of states is shown for *each* type of inner vacancy.

The Ritz-combination principle, according to which the energy of an emission line is the simple difference between two absorption-edge energies, has only qualitative validity; it is not generally valid in detail.

vacancies gives rise to an abnormal configuration of the valence electrons.

Three other phenomena could also be mentioned in this section, all of them of interest in absorption spectra only. These are the interactions of the ejected electron with the other atoms and electrons in a polyatomic system. These interactions—plasma oscillations, and scattering or diffraction—are better discussed later as factors in the transition probabilities, rather than treated as giving rise to separate states.

X-Ray Energy Level Diagram

Figure 7 presents a part of a qualitative energy level diagram which, it is believed, has provisions for the necessary and sufficient complexity to interpret ordinary x-ray spectra. It is designed to avoid the more obvious difficulties encountered with the conventional one-electron-jump type of diagram which is shown in Fig. 1. Figure 7 is drawn for a solid, specifically for an insulator; a similar diagram may be drawn for a metal—the zero energy would appear at or near the bottom of the V (valence) continuum; i.e., near the top of the V_τ band as this band appears in the figure. As compared with Fig. 1, the level arrangement has been inverted; the K group of levels, having the greatest energy, and the energy being positive, appears at the top.

Since the range of spectral energies of primary interest in solid-state work is about the same for all x-ray spectral regions, the diagram is drawn on a linear energy scale. For purposes of illustration, scale sections for only the K, L_{III} and V (valence) groups of levels are shown in Fig. 7, and only a few (four) of the *many* possible discrete levels in each case (except the V case, see later) are shown. A logarithmic scale would again be convenient if the relative energies of many groups of levels of a solid containing an atom of high atomic number were to be indicated coherently in the diagram, but, then, to show the multiplicity of discrete levels due to the splitting, a magnified insert of each group of high-energy levels (e.g., the K group or the L_{III} group) would be required.

The actual separations of the peak positions, the degree of overlap of adjacent states, as well as the relative prominence of the individual members of a split group, are generally not yet known. The diagram of Fig. 7 must be recognized as being merely qualitative in these regards. And a sensible terminology for the various levels in each group must await further knowledge.

Both emission and absorption transitions are indicated. The energy of the system in emission does not include the energy of the ejected electron if it has escaped completely from the system, i.e., escaped from the local region of interest in the solid. But if it has not escaped, it and its energy must be included as part of the system. Although, in principle, all of the discrete states and a few of the continuum states are potentially significant as initial states in emission, it turns out that, in ordinary practice with a conventional x-ray tube, the probability is generally high for a transition from only *one* initial state of a split group (i.e., from only one K state or from only one L_{III} state, etc.). However, this high probability for the involvement of only *one* state in each of the two complex groups of states in emission is *not* correct when the state in question is of low energy, i.e., of energy less than perhaps 20 ev. This point arises again in reference 70.

Let the K level marked "K_τ" refer to the electron configuration in which, conventionally, the K electron has been removed to "infinity" in the solid and left there with zero kinetic energy. (The condition of zero kinetic energy is of interest in absorption transitions only.) It is suggested that this is the significant K level in the ordinary K-emission spectra, whatever is the configuration of electrons appropriate to this level. ("Ordinary K emission spectra" refers to spectra from K states of single inner-vacancy produced by high-energy cathode electron bombardment as in a conventional x-ray tube.) In this sense, we may call K_τ the "unexcited K level."[46] All of the other discrete K levels refer to K states in which a different electron configuration obtains. This different configuration, as already pointed out, may be one in which the ejected K electron is still bound, albeit lightly, to the positive hole it left behind, and/or in which a different number of outer

[46] The terminology here becomes a trifle subtle: This particular K state, K_τ, is an ionization state by definition; all other discrete K states are excitation states, some being excited (but not ionized) on each of two or more counts and some being excited ionization states.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

electrons failed to jump down closer to the atomic nucleus as the electric field changed upon the production of, or subsequent to the production of, the K-electron vacancy.[47] The continuum of levels indicated by the diagonal lines above K_τ in Fig. 7 represent the theoretical normal density-of-states in the so-called unoccupied continuum, but also in this region are some unmarked discrete levels.[48] In the case of a solid, "normal" refers to the theoretical model in which there is no perturbation due to the inner vacancy; any perturbation effects are presumed to be included in the discrete levels or bands. The theoretical normal continuum has maxima and minima in it, and these are supposed to correspond to "Kronig structure" in x-ray absorption spectra, as discussed later.

One measurement of the difference between the energy of K_τ and about the lowest of the bound-ejected-electron excitation states is afforded by a comparison of the energy required to produce the K state when (1) a large excess energy is available in the process and (2) when no excess energy is available. In the case of metallic copper, the former energy is 8985.5 ev[59] and the latter is 8980.7 ev[1].

Again, what has been said about K states also applies in general to states in which the vacancy is in some other shell or orbital, including those states in which the vacancy is in one of the valence orbitals. But the valence states are significantly different.

By definition, the ordinary x-ray energy level diagram gives the energy of the system when one and only one electron has been lost from the shell or orbital designated in the level terminology. This means, in a strict sense, that the *normal* unperturbed valence band of the solid should not be included in the diagram. But the assumption is often made that a vacancy in a normal valence orbital does not significantly perturb the pertinent region of the solid, and hence that it does not significantly perturb the energy or shape of the normal valence band. So, with this assumption, we include the normal valence band in the diagram. This is the band labeled "V_τ" for which the ejected electron is at "infinity" with zero kinetic energy.

By identifying V_τ with the normal valence band, we have pre-empted excitation states, or bands of states, of the valence-electron-configuration type. Unless the perturbation, which is assumed in the previous paragraph to be negligible, is not negligible (or does not remain negligible as the number of valence vacancies increases slightly), all excitation bands are degenerate and are the same as V_τ. The other V bands indicated in Fig. 7—V_σ, V_ν—are representative of the various kinds of bound-ejected-electron excitation bands.

The width and shape of the valence band refer to the density of states regardless of admixed wave-function symmetries and regardless of the probability of position of the positive hole in the band. But, in particular, it may be that in some cases in absorption the position probability is sharply peaked near the top of the band, whereas in emission the vacancy may be more likely near the bottom.[49,30,31] Even if it were preferentially at the bottom initially, e.g., after the jump of a valence excitation electron (from an excitation orbital inside the normal valence orbitals) to an inner vacancy, the hole may tend to "move up" in the band as a consequence of a second valence electron jump, in this instance within the band. Such a double-jump radiative transition is perhaps not very probable.[50] However, it is conceivable that the effective position of the final valence hole may depend upon the particular solid and in some instances may even be time dependent. If such differences in the position probability are real, they are not properly a feature of the diagram *per se*, but rather are part of the transition probability correction (h) as listed earlier.

Another time-dependency feature in the relative transition probabilities has to do with radiationless (Auger) transitions or with the interaction with lattice vibrations, i.e., with phonons. If an inner vacancy state lives long enough for significant Auger or phonon interaction to take place, readjustment of the configuration of the outer electrons may occur in such a way as (a) to quench or partially quench some of the excitation states in favor of others or of the unexcited ionization state, and (b) to alter the wave-function symmetry of an outer bound electron and thus to introduce somewhat different excitation states. These changes in relative transition probabilities are in addition to the Auger broadening of states owing to reduction of lifetime, and also to broadening due directly to lattice vibrations as discussed later.

Of course, if the readjustment of the outer electrons is not part of the radiative transition, the angular momentum of the system must be conserved in the radiative transition itself. The angular momentum, which in absorption includes the angular momentum of any ejected electron *no matter where it is*, must be

---

[47] For example, the proposed states involved in Mn$K\beta'$ emission are among those symbolically represented as ρ and σ states, i.e., $K_{\rho,\sigma} \rightarrow (M_{II,III})_{\rho,\sigma}$, although these states are not intended to be interpretated so specifically in this generalized diagram.

[48] In a more general sense, when our concepts include the interactions among *all* the electrons and nuclei in the solid, no distinction *in kind* need be made between the continuum states and the discrete states either below or in the continuum.

[49] It is common practice in ultraviolet absorption spectroscopy with insulators to depict the first valence excitation band (called an "exciton") as a narrow level instead of as a band. This practice presumes that the position probability of the positive hole in the band is sharply peaked, usually taken as at the top of the band because of the calculated theoretical shapes of the bands or energy *vs* wave vector (one-electron model calculations) and by an application of the principle of conservation of the momentum of the system.

[50] However, Landsberg[32] calculated that the up-movement of the valence positive hole by an Auger or radiationless process in certain metals might take place in less than 10^{-16} sec, a very short time. If the desire for the hole to move up is this great, the double-jump radiative transition may not be so improbable.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

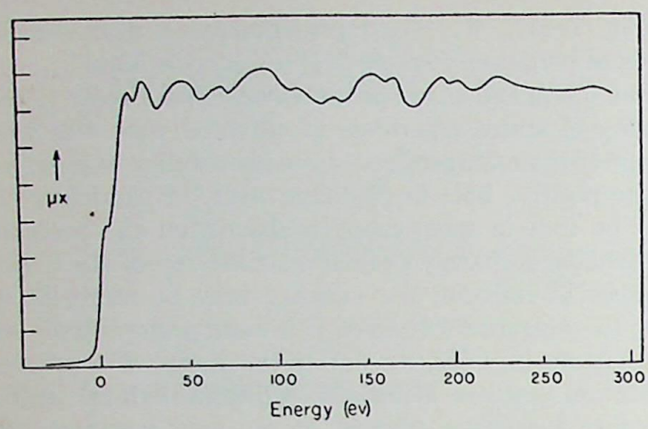


FIG. 8. Extended K-absorption structure (uncorrected) for metallic copper.

reckoned with respect to a reference axis passing through the local atom.

In the case of atoms in an electrically polarized environment, e.g., in a complex ion in a solution or in a solid, many or all of the orbitals of the bound-ejected-electron may be significantly compressed in radii and constrained within a region bounded by a strong localized potential barrier (perhaps by an electrical double-layer). For such atoms, the close-in structure in the absorption spectra are likewise appreciably altered and made much less sensitive to the more remote atomic environment.

Finally, it must be recognized that the diagram of Fig. 7 refers to ordinary x-ray states, and that different but similar diagrams must be drawn for x-ray states involving two (or more) inner electron vacancies. Of course, both emission (satellite) and absorption transitions may be indicated on such diagrams for the multiple-inner-vacancies.

Absorption Spectra: Extended Structure

To recapitulate, the important x-ray states in absorption spectra are (a) bound-ejected-electron excitation states, (b) valence-electron-configuration excitation states, (c) multiple-inner-vacancy states, and (d) the continuum states, including the effects of plasma oscillations and of scattering or diffraction of the ejected electron in the x-ray system. But, generally, the unexcited ionization state, *viz.*, K_τ or L_τ, etc., is *not* significantly involved in absorption spectra close (within say +20 ev) to an absorption discontinuity, notwithstanding widespread elementary notions of the contrary. For the final state in absorption, the geometrical extent of the x-ray system always includes the ejected electron and is therefore usually much larger than is the extent for a state in emission. All the high-up continuum levels are in the diagram for absorption interests only. For emission all these continuum levels become simply K_τ or L_τ, etc., since for them the ejected electron is reckoned simply as having escaped from the system; the energy with which it escaped is of no interest in the energy of the state.

For a monatomic gaseous absorber (e.g., see Fig. 2) *all* K-excitation states of the bound-ejected-electron type appear as levels below the K-ionization level when these levels are presented in a diagram of the type of Fig. 7. In a solid, the K_τ level is proposed as the closest equivalent to the atomic K ionization level. In Fig. 7, one excitation level is drawn above K_τ (or L_τ, etc.); this is partly arbitrary and partly due to the fact that, in a solid, K_τ does not have the simple "series limit" meaning that it has in a monatomic gaseous atom.[51]

In absorption, the relative prominence of the different types of excitation states, compared with the continuum of states, depends upon the particular absorbing material and the particular inner vacancy. For example, for crystalline potassium chloride, discussed later, the excitation states dominate in the region near the "main" K-absorption edge, whereas for metals the excitation states are less prominent. Extended absorption structure, due to scattering or diffraction of the ejected electron, is observed for all polyatomic systems,[52] and, for solids, it seems to be more prominent for metals than for insulators.

Figure 8 illustrates the extended structure typical of crystalline metals. Numerous irregular undulations are observed on the high-energy side of the apparent edge. Although very important in the first 20 ev, the excitation states cannot be responsible for much of the structure 50 or more ev beyond this region. The observed irregularities beyond about 20 ev are commonly referred to as Kronig structure in honor of the first investigator to propose their interpretation (in 1931).[26] According to Kronig, zero absorption is predicted if the ejected electron has such a combination of speed and direction in the crystal lattice that it suffers Bragg reflection[53] from the lattice planes; and a maximum in absorption is predicted if the ejected electron does not have such a combination and does not undergo Bragg reflection. The predicted abrupt changes in absorption become smoothed into mere undulations when various directions in the crystal lattice are simultaneously taken into account. In the band theory of solids, as it developed subsequent to about 1928, a "forbidden" region of energy corresponds to just those values of vector momenta of the electron that allow Bragg reflection (the boundary of a Brillouin zone). An allowed energy

---

[51] There is a strong tendency in the literature for the assumption that a similar or identical series-limit meaning holds in all kinds of absorption spectra. See reference 1 and G. H. Wannier, Phys. Rev. **52**, 191 (1937). Also, Vainshtein, Barinskii, and Narbutt, Zhur. Eksptl. i Teoret. Fiz. **23**, 593 (1952). For a solid, a series-limit for bound ejected-electron excitation states no doubt exists but it is not, in general, coincident with or very near to the bottom of the conventional solid-state unoccupied continuum, nor is it the same limit for different inner vacancies or for different atom species in the solid. This lack of coincidence is due mostly to the respective wave-function extents and symmetries, but partly to the time factors.

[52] H. Peterson, Z. Physik **80**, 258 (1933).

[53] The general scattering or diffraction of the ejected electron becomes Bragg reflection in a crystal if the electron is presumed to be a plane wave in the lattice.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

band is that for which the electron can travel freely through the lattice without reflection. Thus, after the spectra have been corrected for the various effects (a) through (h) listed earlier, and if the entire intensity pattern is considered, instead of merely the positions of the maxima and minima, the observed Kronig structure is just the x-ray absorption view of the theoretical normal density of states in the overlapping unoccupied bands of the solid.

In Kronig theory, the inner vacancy is presumed not to perturb the normal bands in any way. It is probable that the influence of this perturbation is such as to complicate seriously any Kronig interpretation of observed structure within 30 or so ev of the apparent edge, and to be a significant factor as far out as 75 or 100 ev. Also, in Kronig theory, the ejected electron is presumed to be a plane wave, i.e., an electron rather far removed from the vacancy site. Actually, for an x-ray state of short lifetime, the ejected electron does not have time enough to travel very far from the vacancy site (a 50-ev electron of normal effective mass would have time to travel a total distance of only about 4 A in 10^{-16} sec), and the plane wave presumption is perhaps not very good except for the very extended absorption structure.[54]

An extension of Kronig's interpretation has been offered by Hayasi.[55] According to Hayasi, if the Bragg reflection occurs at a Bragg angle near 90°, the ejected electron remains in the vicinity of the inner vacancy as in a type of standing wave pattern. An increase in the absorption probability results because of the greater overlap of the wave functions of the initial and final states. Thus, Hayasi predicts an absorption maximum for some of the energies at which Kronig predicts a minimum, or at least an increased absorption at some of Kronig's minima. In the final state, a "quasi-stationary" state as Hayasi calls it, the ejected electron is, in a sense, bound to the positive hole it left behind, and, in another sense, the electron is in a forbidden energy region in the solid. Insofar as the electron is bound to the positive hole, the final state is one form of what we have called an excitation state. Hayasi, like Kronig, ignores the lattice perturbation due to the inner vacancy. Moreover, for a short-lived inner-vacancy state, the ejected electron does not have time to establish itself as a very respectable standing wave.[56] For comparison with experiment we need more detailed and quantitative calculations on the basis of Hayasi's proposal, with inclusion of the vacancy perturbation, the time-dependencies, and all directions of ejection in the lattice.

Efforts to explain the observed details of absorption curves on the basis of Kronig's interpretation have resulted in some but not complete success. Hayasi's modifications may possibly help. The excitation states will probably account for most of the structure in the close-in regions. But it is quite likely that additional processes are involved in the complete explanation. It has been suggested that a correlation exists between the x-ray absorption structure and the discrete loss of energy of electrons as they penetrate thin solid foils.[57] We may distinguish two types of discrete energy loss. The first is in the production of multiple inner-vacancy states. In any type of absorber, the interaction of the electromagnetic field with the system may result in the ejection of two (or more) electrons from an atom instead of just one electron. One or both ejected electrons may come to rest in an orbital or orbitals still bound to the *two* positive holes left behind. Or the second electron may come from a different atom in the system. In any of these cases, the secondary electron-jump absorbs a discrete amount, or at least a minimum amount, of energy and provides an additional way in which the system can absorb the energy of the incident x-ray photon. (The minimum amount of energy absorbed by the secondary jump is essentially zero only for the special case of an electron at the Fermi surface in a metal.) When the photon energy and system are such that the probability for the secondary jump is reasonably high, the absorption should exhibit an increase.[58] There are many potential secondary electrons in a solid, and the discrete amounts of energy become very numerous when more than one secondary electron and more than one atom are considered. The absorption structure due to the production of multiple-vacancy states may indeed show only one or two maxima and an accompanying broad absorption tail of unresolved components.

The second process for a discrete loss of energy is by the induction of plasma oscillations. The ejected electron as it traverses the crystal interacts with the loosely bound valence electrons in such a way that the electrons oscillate collectively. The energy of the oscillation is taken from the absorbed photon via the kinetic energy of the ejected electron. If more than one mode of oscillation is induced, the electron loses more than one discrete unit of loss energy, and if more than one oscillation of a given mode is induced, the electron loses a multiple of a discrete unit. As in the case of

[54] In a recent paper, Shiraiwa, Ishimura, and Sawada [J. Phys. Soc. Japan **13**, 847 (1958)], discuss a modification of Kronig's theory by allowing the amplitude of the wave of the ejected electron to decrease rapidly with distance from the parent atom.

[55] T. Hayasi, Sci. Repts, Tôhoku Univ. **33**, 123, 183 (1949); **34**, 185 (1951); **36**, 225 (1952).

[56] To the degree that the general position of the Hayasi electron is localized, the uncertainty principle argues further against its being a plane wave with well-defined momentum. The Hayasi reflection at best should be considered as "quasi."

[57] E.g., see Leder, Mendlowitz, and Marton, Phys. Rev. **101**, 1460 (1956); Revs. Modern Phys. **28**, 172 (1956); E. J. Sternglass, Nature **178**, 1387 (1956); D. Pines, Revs. Modern Phys. **28**, 184 (1956); R. H. Ritchie, Phys. Rev. **106**, 874 (1957).

[58] Production of multiple-vacancy states in gaseous absorbers should be about as probable as in, say, ionic crystalline absorbers. Reports indicate very little absorption structure for gases that can be attributed to multiple vacancies. For this reason these states may not play a very significant role in absorption, but the argument is not generally conclusive.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

ejection of a second electron, plasma-oscillation production represents an additional way in which the system can divert the energy of the x-ray photon. An increase in the absorption probability results for those particular photon energies properly tuned.

The energy spectrum of x-ray photo-ejected K electrons from metallic copper has been analyzed magnetically[59] and found to consist of a secondary maximum and a long low-energy tail. It is not yet clear which of the two types of inelastic scattering of electrons is primarily responsible for the observed structure.

As implied earlier, these two discrete-loss processes also take place when no primary inner electron vacancy is produced; in fact, this is normally the situation in the experiments of penetration of electrons through thin solid foils.[57] But the probability with which each process takes place is believed to be increased when the primary inner vacancy is made at the same time. For example, in producing the multiple-vacancy state, the primary electron ejection may, by the abrupt change of the Hamiltonian of the system, trigger the secondary electron jump. This may be especially so in case the secondary electron is a valence electron. (In case the second vacancy is in the valence band, the state would be, according to our terminology, a valence-electron-configuration state.)

In each of the aforementioned possible factors in the interpretation of the extended structure in absorption, a reference "zero energy" is involved. It is usually very difficult or impractical to disentangle this zero energy from the observed structure with the requisite accuracy to make definitive correlations between experiment and theory. To date, only the maxima and minima in the structure have been considered. It would be much more informative if complete absorption patterns were correlated. Then we could better see what absorption remains to be accounted for by additional processes such as discrete loss.

Widths and Lifetimes of X-Ray States

With a sufficiently open energy scale, preferably linear, the width (and the gross shape) of at least one member of each group of levels may be conveniently shown on one side of the diagram of Fig. 7. It should be mentioned that the level width and shape in the case of crystalline solid materials should include the effects of lattice vibrations at a particular temperature. These effects are discussed in some detail later in connection with the interpretations of the spectral curves for potassium chloride. Typically, lattice vibrations at a temperature of about 300°K contribute a few tenths of an ev (or less) to the full width at half-maximum, and contribute a Gaussian factor in the shape. The numerical values of widths indicated in Fig. 7 are about right for chlorine in crystalline potassium chloride if the effects of lattice vibrations are neglected.

The widths of the energy levels corresponding to a few inner-vacancy-states, excluding the effects of lattice vibrations, are given in Table I. These widths, deduced from measurements with the two-crystal type of spectrometer whose resolving power is rather well known, are believed to be essentially atomic values, i.e., unperturbed by (or else corrected for) the physical or chemical environment. Corrections have been made for the instrumental spectral window, for the backgrounds, and for the overlapping lines.

Determination of the lifetime of a state follows from the width according to the uncertainty relation $\Delta t \Delta E \simeq h/2\pi$; a full width at half-maximum of $\Delta E = 1$ ev corresponds to a lifetime of $\Delta t \doteq 6.6 \times 10^{-16}$ sec.

The K width for gaseous argon, 0.5 ± 0.05 ev, obtained from the K-absorption transition $K \rightarrow N_{II,III}$ (or $1s \rightarrow 4p$), is perhaps the most reliable value in the table, and this serves as a sort of anchor point in the determinations of some of the other values.[60] This argon K width is reasonably consistent with the combination of (a) experimental values of the fluoresence yield, and (b) a calculation for the sum of the theoretical radiative transition probabilities.[61] The K and L_{III} widths for gold were taken from the width of the respective absorption edge.[62] This is admittedly not a reliable method for obtaining a level width because of the unknown role of component structure due to the excitation states. However, it happens that both of these gold widths are in fairly good agreement with the calculated theoretical values.[62,63] Furthermore, for the gold L_{III} width, the consistency argument of Richtmyer, Barnes, and Ramberg is reasonably persuasive. (This argument considers the widths of *many* different emission lines.) The same type of argument supports the silver values.[64,65] For every element in the table other than argon and gold, the state widths were obtained from the $K\alpha_1$ emission line. All emitters were

TABLE I. Full widths at half-maximum of x-ray atomic energy levels corresponding to the states indicated (in ev).

State \ Atom	$_{16}$S	$_{18}$A	$_{22}$Ti	$_{25}$Mn	$_{29}$Cu	$_{32}$Ge	$_{36}$Kr	$_{42}$Mo	$_{47}$Ag	$_{79}$Au
K	0.35	0.5	0.8	1.05	1.5	2.0	3.0	5.0	7.5	54
L_{III}	0.1	···	0.25	0.35	0.5	0.8	1.1	1.7	2.2	4.4

[59] Sokolowski, Nordling, and Siegbahn, Arkiv Fysik 12, 301 (1957).

[60] The argon absorption curve is being restudied (with higher resolving power, quartz crystals instead of calcites) to obtain still more reliable data on the state shape and width, and on the relative transition probabilities of the series $1s \rightarrow np$, $n \geq 4$.

[61] D. L. Dexter and W. W. Beeman, Phys. Rev. 81, 456 (1951).

[62] F. K. Richtmyer, Revs. Modern Phys. 9, 391 (1937); Richtmyer, Barnes, and Ramberg, Phys. Rev. 46, 843 (1934).

[63] L. Pincherle, Physica 2, 596 (1935).

[64] L. G. Parratt, Phys. Rev. 54, 99 (1938).

[65] As judged now with benefit of the new correction procedures,[17] a small over-correction for the instrumental spectral window was reported in reference 64 for most of the silver L series lines. The corrections reported in reference 62 for the gold L lines are also rather dubious, but this is not of great importance now.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

solids[39] except krypton which was gaseous.[66] For titanium, manganese, and copper, the $K\alpha_1$ line is markedly asymmetrical and correction for this asymmetry is discussed in the next paragraph. In each case, the corrected line width was divided between the K and L_{III} states, each state being assumed to be a singlet and of Lorentzian shape. In making the division, two criteria were kept in mind: (1) the radiative components of the K widths should be a reasonably simple function of atomic number Z, and (2) the ratio of the width of the K-radiative component divided by the total K width should agree reasonably well for each Z with the experimental value of the fluorescence yield.[67] The K-radiative widths so obtained are listed in Table II. The K fluorescence yields by this method are also listed. It turns out that the dependence of the radiative width on the atomic number is in good accord with the theoretical Z^4 prediction.[68] (The radiative component of the L_{III} width is very small, much less than the uncertainty in the total L_{III} width.)

As stated in the foregoing, the $K\alpha_1$ lines for titanium, manganese, and copper (and, of course, for some other elements not included in the table) are asymmetrical in shape. When the widths and asymmetries of these lines are each plotted *vs* atomic number through the range $Z=16$ to 42, it becomes obvious that the apparent excess width is closely related to the asymmetry.[39] Illustration of the line shape, after correction for everything except asymmetry, is given in Fig. 9. It is proposed that the excess width on the low-energy side is essentially attributable to transitions between excitation states of the valence-electron-configuration type, just as in the case of the β' structure in the Mn$K\beta_{1,3}$ region discussed earlier.[69] Tentative calculations show

TABLE II. K-radiative widths (in ev) and fluorescence yields.

State \ Atom	$_{16}$S	$_{18}$A	$_{22}$Ti	$_{25}$Mn	$_{29}$Cu	$_{32}$Ge	$_{36}$Kr	$_{42}$Mo	$_{47}$Ag	$_{79}$Au
K_{rad}	0.04	0.07	0.2	0.33	0.65	1.0	1.7	3.6	6.0	50
ω_K	0.1	0.14	0.22	0.31	0.43	0.5	0.57	0.72	0.8	0.93

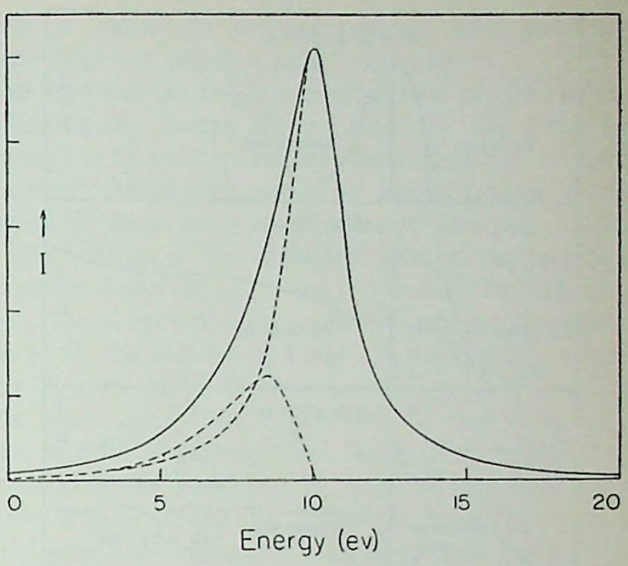


FIG. 9. Asymmetric $K\alpha_1$ "line." The line $K_\tau \rightarrow (L_{III})_\tau$ is presumed to be symmetrical (see text).

that the energy separations between K_τ and the pertinent K excitation states, between $(L_{III})_\tau$ and the pertinent L_{III}-excitation states, and between the $(M_{III})_\tau$ and the pertinent M_{III}-excitation states, are such that now we observe mere asymmetry instead of the plateau region observed in the $\beta_{1,3}$ region. The line from the transition $K_\tau \rightarrow (L_{III})_\tau$ is presumed to be symmetrical and to be obtained by subtracting the mirror image of the high-energy side of the maximum ordinate of the corrected observed curve from the low-energy side. It is the width of this symmetrical component that is used in deducing the K and L_{III} level widths.[70]

In support of this analysis of the asymmetry are the facts that the widths so corrected for asymmetry have a fairly smooth dependence on atomic number over a range of Z that includes both symmetrical and asymmetrical lines, and that the radiative component of the K width varies as Z^4. In support of the contention that these widths are essentially atomic, i.e., relatively free

[66] E. Wilhelmy, Z. Physik **97**, 312 (1935).

[67] Broyles, Thomas, and Haynes, Phys. Rev. **89**, 715 (1953); E. S. Burhop, *The Auger Effect* (Cambridge University Press, London, 1952).

[68] The report by W. H. Zinn [Phys. Rev. **46**, 659 (1934)] of a Z^4 dependence of the *total* widths is based on measurements of the widths of absorption edges that are evidently replete with excitation structure. These measurements are not reliably interpreted in terms of level widths. Typical of this difficulty, Zinn used the "corrected" K-level width for silver as 25 ev, whereas its actual value is near 7.5 ev. However, the Z^4 prediction is expected to be fairly good for reliable total widths when the Auger contribution is small, as for K widths for $Z \gtrsim 45$.

[69] Some of the asymmetry may be due to the other effects mentioned in connection with the $K_{1,3}$ region, *viz.*, unidentified satellites, Compton and Raman scattering, and interaction between the inner L_{III} vacancy and the incomplete $M_{IV,V}$ electron shells. However, these effects are here presumed to be of negligible importance; but whether they are or not is not relevent to the present argument for obtaining the level width. The pertinent assumption is that all such effects distort the line $K_\tau \rightarrow (L_{III})_\tau$ on only one side.

[70] A final caution may be expressed in regards the numerical widths of x-ray energy levels. If, as is supposed, the excitation states are responsible for asymmetry in an emission line, these states may also contribute to the width of the symmetrical component of the line. In other words, the widths as listed in Tables I and II for single energy levels may be a little too great. If the symmetrical part of an apparently single energy level is itself the sum of several closely spaced narrower components, then, even if each component is of Lorentzian shape, the symmetrical component may be expected to be non-Lorentzian, perhaps more nearly Gaussian. Thus, in the question as to whether the width in Tables I and II is about right for a single state or refers to the width of a composite of excitation states, we find two experimental features of interest: (1) the relative magnitude of the width, and (2) the shape of the state. As judged from present information, the K and L_{III} state widths as listed seem to refer in each case to essentially singlet states, but the evidence is not clear in the cases of M, N, etc., states. E.g., the $M_{II,III}$ states of Fig. 6 are wider than the K state for manganese, and the $M_{II,III}$ states may well be Gaussian in shape. These M states may be wide because of (a) a high probability for Auger transitions, (b) phonon interactions (discussed later), or (c) static interactions of neighboring atoms (solid-state broadening).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

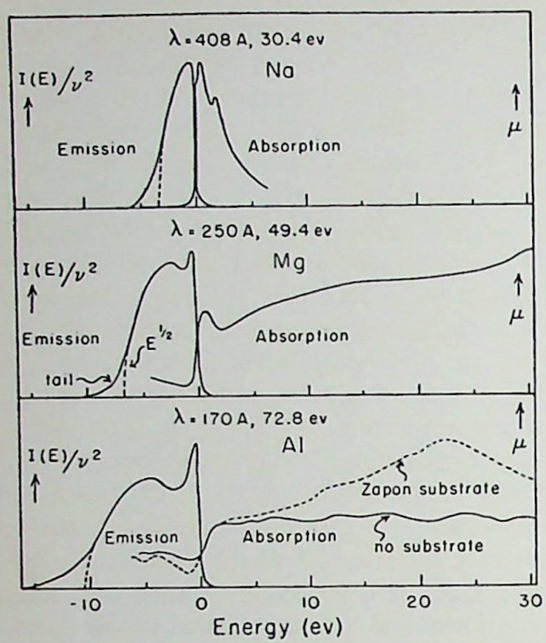


FIG. 10. L_{III} emission and absorption spectra for metallic sodium, magnesium, and aluminum. The same energy-interval scale is used for all the curves; zero is taken as at the estimated position of the Fermi energy. The curves are taken from various authors.[71]

of solid-state effects, is the fact that the gases argon and krypton are included in this atomic number dependence without distinction from the solids.

V. EXPERIMENTAL CURVES AND INTERPRETATIONS

The work in this subject is far too extensive and intensive to be summarized adequately in a few pages. In the preceding sections, only a few topics were selected for discussion; and, in this section, only a few additional curves are selected as illustrative of the experimental results and interpretations.

The interpretations are not clean-cut, in general. For the most part, the corrections listed earlier have not been made to the observed spectral curves; and, in particular, the recognition and use of excitation states is relatively new. A few satisfactory features can be pointed out, but the reader is warned to expect some confusion. An interpretational "break-through," however, may not be far away.

Sodium, Magnesium, and Aluminum

First, we mention briefly the L_{III} spectra of metallic sodium, magnesium, and aluminum. These spectral curves, reproduced in Fig. 10 from various authors,[71] were obtained with sufficiently high resolving power that no correction for this effect need be made. Also, great care was taken (except possibly in the case of sodium absorption) to translate the observed blackening of the photographic plates to an x-ray intensity (or absorption) scale.

There are several features of interest in these curves, but to see them in perspective we first recall the general electronic aspect of these solids. The electron configuration of atomic sodium, atomic number 11, is $1s^2 2s^2 2p^6 3s$ in which the $3s$ electron is nominally in the solid-state valence band, i.e., in the occupied part of the conduction band. For magnesium, atomic number 12, we find two nominal $3s$ electrons; and for aluminum, atomic number 13, we add one more electron, this time in the nominal $3p$ band. It is the combination of $3s$ and $3p$ bands from which the electron comes to fill the inner L_{III} vacancy in the L_{III} emission transition.

The dipole selection rule dictates that only an electron having wave-function symmetry of s type or d type, as viewed from the L_{III} vacancy, may fall radiatively into the L_{III} vacancy (p-type symmetry). Hence, to the extent that the nominal $3p$ band is involved, there must be admixture of s- and/or d-type symmetries in it.[72,73] The selection rule actually applies to the wave functions of the *states*, but, as pointed out earlier, interpretation of this rule in terms of the one-electron model suffices if the primary jumping electron is the only electron in the entire system whose angular momentum changes.

In the electron configuration for sodium, the nominal $3s$ band is half occupied; hence, sodium is a good conductor. The $3s$ band in magnesium is fully occupied and, hence, since magnesium is a good conductor, the $3s$ and $3p$ bands must overlap. Aluminum is a metal because the $3p$ band is only partially occupied. Magnesium has a different crystal structure than sodium and aluminum, but theoretical calculations suggest that the overlap is similar for all three of these metals, the principal difference being in the measure of occupancy. The so-called reduced width of each emission band (i.e., the base width corrected for the "tail" as indicated by the respective dashed curve in the figure) is conventionally taken as a measure of the occupancy of the $3s$, $3p$ bands. The energy widths of the $3s$, $3p$ bands, as obtained directly from the measured bands, are 3.05, 6.4 and 10.6 ev, respectively, and this is interpreted as showing the increase of electron occupancy as we go from sodium to aluminum. It is possible that the role of excitation states is not important in the width and shape of these curves, but, on the other hand, it is conceivable that the valence-electron-configuration type of excitation states do influence the curves in some manner similar to that discussed earlier for the Mn$K\beta_{1,3}$ region. If the excitation states are important, then only the relative magnitudes of the observed widths may be significant in showing relative

[71] The emission curves are taken from W. M. Cady and D. H. Tomboulian, Phys. Rev. 59, 381 (1941); the absorption curves from H. M. O'Bryan, Phys. Rev. 57, 557, 995 (1940), J. R. Townsend, Phys. Rev. 92, 556 (1953), and D. H. Tomboulian and E. M. Pell, Phys. Rev. 83, 1196 (1951).

[72] Although for quadrupole radiation, p to p transitions are allowed, this radiation is very feeble compared with dipole radiation.

[73] It is also important to note that, at least in hydrogenic theory, the transition probability for p to s is much less than (of the order of one-tenth) for p to d.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

electron occupancy, not the numerical magnitudes themselves.

The narrow spike observed at the high-energy side of the emission band for both magnesium and aluminum is conventionally explained as a sharp rise in the density of $3p$ states just before the top of the $3s$ band is reached.[74] The spike is markedly higher for aluminum. This explanation requires that there be a *very* large admixture of *s*- and/or *d*-type electrons in the bottom part of the $3p$-electronic band. The theoretical density of states, according to the one-electron model, with complete disregard for the wave-function symmetries, i.e., for the transition probabilities, is indeed much like the observed emission curves for all three metals with the Fermi energy appropriately placed. But the calculations have not yet succeeded in giving the degree of admixture of wave-function symmetries of the electrons, or of indicating the role of the possible excitation states. These facts should be taken into account before the explanation of the spike or of the general shape of the curve is considered to be satisfactory.

Next we note that the edge of each of the observed emission bands at the Fermi energy is abrupt, as expected from the degenerate-electron gas theory if the inner electron vacancy causes no perturbation. Indeed, when the small corrections are made for the spectral window of the spectrograph and for the width of the initial L_{III} state, the remaining width of the observed emission edge has been attributed to the effect of the thermal motion of the electrons of the emitting anode.[75,76] Reckoned as between the 5% and 95% points on the edge, the temperature width as deduced from the Fermi distribution formula is 0.06 ev at 110°K, and is 0.45 ev at 670°K. The experimental widths at each of several temperatures in this range are found to be in agreement with these theoretically predicted values.[77,75] However, these three metals are most unusual, apparently unique, in that their L_{III} spectra exhibit this abrupt emission edge, an edge that corresponds so nicely to the Fermi surface predicted theoretically for unperturbed metals. This prediction for metals is based on an assumption that the Fermi surface does not touch a zone boundary.[78]

The long tail on the low-energy side of each of these emission bands (some 8% or more of the total band intensity) is at least partially accounted for, according to Skinner[75] and Landsberg,[32] by Auger transitions that occur as the final-state positive hole "bubbles up" from near the bottom of the occupied part of the valence or conduction band of electronic "levels." In this explanation, the Auger process greatly reduces the lifetime of each of the low-lying levels in the band, and the width of each level is correspondingly increased; the tail is thus simply the lower side of the symmetrical spread of each low-lying level. According to this Auger explanation, the intensity in the tail appears at the expense of intensity in the rest of the band, particularly in the lower half. Hence, if a large part of the tail were to be thus explained, the corresponding redistribution of intensity in the band proper would somewhat invalidate the popular comparison of the reduced band shape with the free-electron parabolic relation $I/(\nu)^2$ *vs* $E^{\frac{1}{2}}$ (or *vs* $E^{\frac{3}{2}}$ in the case of *K*-band emission). Other proposals to account for this tail, or part of it, have also been made.[79]

In Seitz's proposal[79] the tail may be due to a series of occupied excitation orbitals just below the valence band. Accordingly, a photon of slightly lower energy is emitted if the primary-jumping electron comes from an occupied excitation orbital instead of from the normal valence band. In these two possible types of transitions, the initial states of the system are presumably the same, and the final state in *each* case is one in which the positive hole is in the valence band. Hence, the proposal must be essentially that the energy position

---

[74] For sodium, there is no evidence in the curve of Fig. 10 for the $3s$, $3p$ band overlap according to this interpretation, but a suggestion of such overlap is seen in Skinner's curves [H. W. B. Skinner, Phil. Trans. Roy Soc. London **A239**, 95 (1940)]. Also, this suggestion appears in a recent work by Fisher, Crisp, and Williams (private communication).

[75] H. W. B. Skinner, Phil. Trans. Roy. Soc. London **A239**, 95 (1940).

[76] This is one of the very few cases in the long wavelength region in which correction for either type of resolving power has been attempted.

[77] Incidentally, it is a truly remarkable experimental accomplishment to know the temperature of the local regions of the anode emitting the x-rays, and also to carry out the necessary photographic procedures across such a steep edge to obtain a *reliable* intensity curve.

This interpretation of the edgewidth does not take into account the effects of lattice vibrations, discussed later in connection with the potassium chloride curves, nor of self-absorption in the x-ray tube anode.

[78] It is often said that lithium, sodium, and potassium are our most nearly ideal metals, ideal in the sense of meeting some of the conditions assumed in a theoretical treatment. Lithium *K* emission does not show an abrupt edge [D. E. Bedo and D. H. Tomboulian, Phys. Rev. **109**, 35 (1958)], and this may possibly be accounted for, according to Dr. V. Heine (private communication), if the Fermi surface touches a zone boundary. Potassium *K* emission, although recorded [E. L. Jossem and L. G. Parratt, Phys. Rev. **98**, 1151 (1955)], awaits corrections (especially for the resolving power due to the *K* state, for self-absorption and for the background due to the $K\beta_1$ line) before the answer can be definite. There is a suggestion of the type of distortion observed with lithium. Lack of abruptness in these cases may possibly be attributed, in addition to the distortion due the Fermi surface touching the zone boundary, to the role of excitation states or to the nominal *s* and *d* types of symmetry in the conduction band. Since the final state in *K* emission must have *p*-type symmetry, the experimental curves tell us the actual (admixed) *p*-type distribution only.

Dr. S. E. Williams, University of Western Australia, reports (private communication) a new observation that the L_{III} emission of potassium metal may also be abrupt, having a width of "0.22 ev including the instrumental width at 1 A resolution." He estimates that his potassium target was at a temperature near its melting point.

R. H. Kingston [Phys. Rev. **84**, 944 (1951)] reports the *M*-emission edge of potassium metal to be rather sharp also.

[79] F. Seitz, *Modern Theory of Solids* (McGraw-Hill Book Company, Inc., New York, 1940), pp. 436–439; Cady and Tomboulian[71]; S. Raimes, Phil. Mag. **45**, 727 (1954); and D. Pines, *Solid State Physics* (Academic Press, Inc., New York, 1955), Vol. 1, pp. 367–450.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

of the positive hole is lower in the band when the jump occurs from one of the excitation orbitals than it is when the jump occurs from the normal band.[80] This sort of argument is also discussed (with an indefinite conclusion) in the case of K emission of potassium chloride by Parratt and Jossem[31] where the excitation spectra are presumed to be more intense than the normal band spectra. [As to the expected relative intensity of these two types of spectra, note that the transition probability (independent of relative quantum weights) in each case is proportional to the degree of overlap of the wave functions of the initial and final states, and that the overlap is greater for the excitation spectra simply because the excitation orbital is closer (more tightly bound) to the nucleus.]

If Seitz's proposal were interpreted as implying different initial and final states for the two different radiative transitions, as different valence-electron-configuration excitation states, the explanation of the tail would be essentially the same as is discussed earlier for the $K\beta'$ "tail" in the Mn$K\beta_{1,3}$ region of Fig. 6, and suggested also for the $K\beta''$ satellite accompanying the $K\beta_{2,5}$ lines and for the asymmetry of the $K\alpha_{1,2}$ lines. Indeed, with this interpretation, the significance of the "reduced width" scheme by which the tail is separated from the main emission band, becomes artificial, and we are really talking about the shape of the emission as a whole.

It is conceivable that, if the Skinner-Landsberg explanation of the low-energy tail is even qualitatively correct, the symmetrical Auger-broadening of the states in the band must be sufficient also to produce at least some of the small but definite tail observed on the high-energy side of the band, even though the states in the uppermost part of the band are themselves negligibly broadened. However, Skinner has proposed another explanation,[75] *viz.*, that this high-energy tail is due to the "satellite" band $L_{III}M \rightarrow MM$, i.e., to the transition between an initial state containing both an L_{III} and an M vacancy and a final state containing two M vacancies. According to the usual satellite argument for states of double (or multiple) *inner* vacancy, the unscreened or residual electrical field, which is present as a consequence of the extra inner vacancy, produces a little extra energy in both the initial and final states in such respective amounts that the emitted satellite photon has a slightly greater energy. Note, however, that the extra M vacancy in the case of any of the present three metals is a vacancy in the valence or occupied part of the $3s$, $3p$ electronic bands, and we have presumed above that if the vacancy is in the normal valence band it does not alter the energy of this band. Such a vacancy in the normal valence band also does not alter the energy of the initial L_{III}-vacancy state either. If these presumptions are valid, the $L_{III}M \rightarrow MM$ band is negligibly shifted and is indistinguishable in shape from the parent $L_{III} \rightarrow M$ band.

But, if the extra vacancy in the initial state is in an M-excitation orbital, then, according to our terminology, the state is called an excitation state of the valence-electron-configuration type, and the energy of the state and of the emitted photon is shifted. By the usual satellite argument, the satellite occurs on the high-energy side of the parent line or band, whereas we have discussed the excitation spectra of this type as shifted to the low-energy side. In the latter case, we have presumed that the normal or unexcited $(L_{III})_\tau$ and V_τ states of ionization in the energy level diagram are those for which there exists one (or more) vacancies in the valence excitation orbitals. When less than the normal number of vacancies is found here, the state is an excitation state, and photon emission therefrom is of lower energy than is the normal line or band. In other words, in terms of the earlier discussion of the manganese K-emission spectra, if the definition of a satellite state were to be extended to include a valence orbital vacancy as the second vacancy (instead of another *inner* vacancy), the Mn$K\beta_{1,3}$ lines would be satellites of the $K\beta'$ parent line or lines, and most of the $K\alpha_1$ line would be a satellite of the low-energy "asymmetrical" component. Or, in terms of the present L_{III} spectra, if the excitation orbitals play a role similar to that proposed for manganese, the observed high-energy spike with its abrupt edge may be a satellite of the low-energy part of the emission. It seems better to retain the definition of a satellite state as given earlier.

The relative intensity of the $L_{III}M \rightarrow MM$ band, as proposed by Skinner without excitation orbitals and regardless of its energy position, would be seriously reduced (as suggested by Skinner) by the rapid transfer of the M vacancy in the initial state from the atom in question to a neighbor and then out of the x-ray system. The time for such a transfer depends upon the position of the M vacancy in the valence band, being probably much less than the L_{III} lifetime if the vacancy is at or very near the top of the band. One Auger transition, $L_{II} \rightarrow L_{III}M$, always places the extra M vacancy very near the top; the $L_{II}-L_{III}$ energy difference is very small, e.g., it is 0.3 ev in the case of magnesium,[2] and therefore the M vacancy thus produced occurs within 0.3 ev of the Fermi surface. But there are other

[80] The discussion given by Seitz[79] is in terms of the one-electron-jump type of "energy level" diagram. He discusses further whether or not the normal valence band "broadens" as a consequence of the inner electron vacancy enough to "absorb" the excitation levels and thereby to obliterate the excitation spectra. (This is presumably not a question of different energies of the initial state; the initial state is presumed to be of the same energy regardless of whether an excitation electron or a normal valence electron jumps to fill the inner vacancy.) The significant questions are where in the band is the final positive hole, and what is the energy of the final state. The answers are *not* given by the one-electron-jump diagram since the excitation level (and any "broadening" of the valence band proper) ceases to exist with the filling of the inner electron vacancy. This is a typical example, of which there are many in the literature, of the difficulties in interpretation with the one-electron-jump type of diagram in solid-state spectroscopy.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

Auger transitions that more frequently place it lower in the band.

Instead of the extra vacancy being in the M band, it may in some cases be in the L_{II} or L_{III} shell, as produced by the Auger transition $K \rightarrow LL$. In this event, the satellites $L_{II}L \rightarrow ML$ or $L_{III}L \rightarrow ML$ are shifted well beyond the Fermi energy and are observed[75] completely separated from the parent band. In shape, these satellites are a small replica of the parent $L_{II,\ III} \rightarrow M$.

The probability for Auger decay of the L_I state is especially very great, and *no* L_I-emission band for any of these metals has ever been reported. Also, the Auger transitions $L_{II} \rightarrow L_{III}M$ and $L_{II} \rightarrow MM$, which "rob" intensity from the L_{II}-emission band, cause the L_{II} band to have only about 1/20 of the intensity of the L_{III} emission for Mg, whereas, on the basis of quantum weights, we would expect an intensity ratio of $\frac{1}{2}$.

As mentioned in the foregoing, some fraction of the intensity thus "robbed" from the L_{II} emission, and likewise some of the intensity "robbed" from the K and L_I emission, reappears as part of the $L_{III}(M)^n \rightarrow M(M)^n$ band where $n \geq 1$; in fact, most of the apparent $L_{III} \rightarrow M$ band is of emission for which one or more extra M vacancies of the normal valence type are present. There is no point in calling most of this emission a satellite or satellites.

Now we turn to the absorption spectra of these three metals. The first point to be made is that the high-energy edge in emission and the low-energy edge in absorption agree in position (at the Fermi energy) as may be theoretically expected of a good metal in the absence of excitation structure. Such agreement is sometimes found for those good metals for which the proper edges can be unambiguously identified, but the proper identification is usually very difficult or impractical. In fact, because of the prospective role of excitation states, the agreement in the present uncorrected experimental curves is perhaps better than we should expect.

Concerning this agreement *vs* disagreement of the emission and absorption edges at the Fermi energy, two effects should be noted in addition to the role of excitation states. First, if the lifetime of the initial state in emission is rather long, significant phonon-interactions may take place and shift the energy of this state before the emission transition occurs. The energy shift is expected to be such as to increase the high-energy side of the emission and, hence, produce disagreement in the positions of the respective edges. Second, if the lifetime of the initial state allows relaxation of the crystal lattice (to relieve the stress set up by the inner vacancy), the energy of both the initial and final states are thereby changed, usually, but not always, reduced.[81] The energy of the inner state is expected to be changed more than that of the outer state, and, hence, such relaxation would also produce disagreement in the respective edges. The lifetime of the MgL_{III} state is probably greater than 10^{-13} sec (Skinner reports[75] the width at 5% maximum to be 0.02 ev), and this is approaching the time required for significant phonon interactions. The lattice relaxation time is a little longer, about 10^{-12} or 10^{-11} sec. These phonon and relaxation effects are not expected to be important when the initial state lifetime is less than about 10^{-13} sec.[82]

There seems to be no obvious explanation of the observed shapes of these three absorption curves beyond the edges. From the conventional interpretation of the emission spectra in terms of the overlapping $3s$, $3p$ bands, we would expect sodium to show strong absorption corresponding to the high-energy emission for magnesium and aluminum. Perhaps it does, but the evidence suggests the presence of additional absorption structure in the range 0 to 5 ev. The magnesium absorption spike, at less than 1 ev from the Fermi energy, seems to be narrower than would be expected on the argument of the aluminum and the $3p$ band, and the aluminum absorption shows no clear maximum in the range 0 to 5 ev.

The fact of *two different* absorption curves in the case of aluminum raises further questions. One of these curves was recorded with the aluminum condensed by evaporation technique onto a Zapon substate; in the other case the aluminum foil was self-supporting. The vacuum-deposited specimen shows a relatively strong broad absorption band beyond about 5 ev above the Fermi energy. There is reason to believe[71] that this strong absorption occurs in a metal-substrate interface region roughly 500 A thick. It is possible that some metal atoms were more or less deeply embedded in the material of the substrate, and that some outer electrons were stripped off the metal atoms in the embedding process. Thus, the neighborhood environment of the boundary aluminum atoms or ions may be quite different than in bulk aluminum metal. There is possibly established, therefore, a wide variety of excitation and ionization states that combine in effect to give the strong but very broad absorption. This type of strong absorption is observed also for celluloid and polystyrene substrates, and indeed for other metals than aluminum.

In general, the interpretation of these spectra, either emission or absorption, is not as yet complete. In this section, we make no effort to present an exhaustive discussion nor a final interpretation. Our purpose is to

[81] The dependence of state energy on the lattice spacing is mentioned later in connection with temperature effects in potassium chloride.

[82] Kostarev [A. I. Kostarev, Zhur. Eksptl. i Teoret. Fiz. 22, 628 (1952); or see M. A. Blokhin, Bull. Acad. Sci. USSR, Phys. Ser. 20, 701 (1956), English translation] proposed that because of lattice relaxation, *every* valence emission curve should be shifted to higher energies in order better to match the edge positions. Kostarev, however, does not discuss the fact that many x-ray states have a lifetime much less than the relaxation time; for these short-lived states, his arbitrary shift would be out of order.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

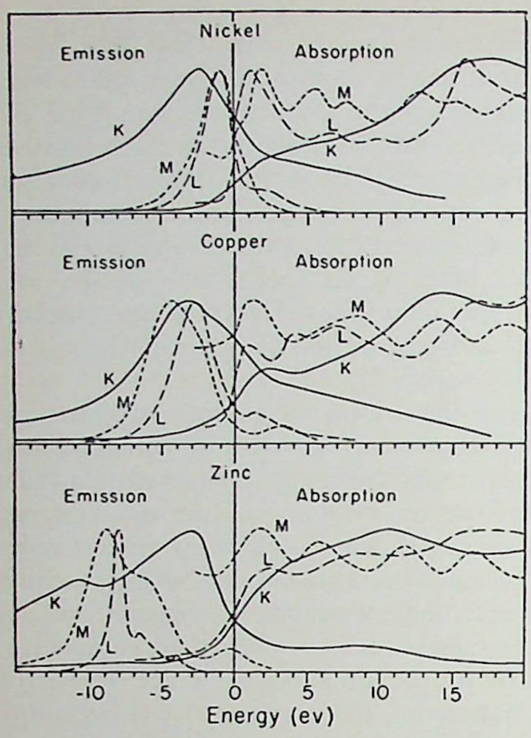


FIG. 11. K, L_{III}, and M_{III} spectra for metallic nickel, copper, and zinc. The curves are taken from various authors.[83]

illustrate typical features of interpretation, and to this end we choose different spectra to emphasize different features.

Nickel, Copper, and Zinc

Figure 11 shows a composite of the K, L_{III}, and M_{III} spectra for three more metals, *viz.* nickel, copper, and zinc.[83] These curves are also normalized and plotted on the same energy-interval scale to facilitate comparison of shapes. The zero of the energy scale is arbitrarily taken in each case as at about the first inflection point of the uncorrected absorption curve[84], the Fermi energy in a good metal is expected to be somewhere near this position. This presumes, as in the case of the sodium, magnesium, and aluminum, (1) sufficient admixture of wave-function symmetries that the different inner vacancies will not, by the selection rule, affect the position of this inflection point, and (2) that excitation states are of negligible importance in this part of these absorption spectra. Neither of these presumptions is generally valid, but there appears now to be no better zero-position criterion. Also, it should be realized that just what is called the first inflection point often depends upon the instrumental resolving power and the precision of the intensity measurements.

The M emission and absorption, as directly observed, consists of overlapping M_{II} and M_{III} bands, but only the M_{III} bands are shown in Fig. 11. In making the necessary resolution, we have taken the energy separation between the M_{II} and M_{III} bands to be small, about 2.5, 2.2, and 2.0 ev for Zn, Cu, and Ni, respectively, in both emission and absorption. Further, we have assumed the relative intensities of the two absorption bands to be 2 to 1 on the basis of quantum weights. In emission, however, the Auger transition $M_{II} \rightarrow M_{III}V$, where V refers to the valence band (admixed $3d$ and $4s$ bands), probably increases this ratio from 2 to 3 or more (and very slightly broadens the M_{II} band); for present purposes, the M_{III}-emission bands in Fig. 11 are deduced from the observed curves with the presumption that the intensity ratio is about 3:1.

None of the experimental curves has been corrected for the corrections (b), (c), (e), (f), (g), and (h) as listed earlier. Although a correction (d) for background intensity has been made in each case by the respective investigators, this correction is often arbitrary. This arbitrariness is probably most pronounced when the background is either changing rapidly with energy (as in the K-emission curves because of the tail of the β_1 line) or is large and somewhat uncertain (as in the M_{III} emission). It is entirely possible that the so-called base width of the K-emission bands in Fig. 11 are too great by 50%, and of the M_{III} bands too narrow by 50%.[85] If the M_{III} bands are as wide and complex as is reported by Bedo and Tomboulian,[85] it seems likely that, in accounting for these features, we must assign a prominent role to the excitation states.

The correction for both kinds of resolving power is large in the K spectra, intermediate in the L spectra, and probably small or negligible in the M spectra. The correction for self-absorption by the anode of the x-ray tube has recently been reported to be very large, at least in certain K-emission spectra.[40] To the extent that an emission band or line, after having been corrected for the instrumental resolving power, straddles an absorption edge, the effect of self-absorption is markedly asymmetrical and causes the emission curve to appear too low on the high-energy side. The self-absorption correction is no doubt important, especially in K spectra for which the jump-ratio at the absorption edge is great.[86] The intensity of the high-energy

[83] Curves for this figure were taken as follows: K emission from J. A. Bearden and C. H. Shaw, Phys. Rev. 48, 18 (1935); K absorption from W. W. Beeman and H. Friedman, Phys. Rev. 56, 392 (1939); L emission from Spielberg, Soules, and Shaw, to be published; L absorption from A. Sandström, Nova Acta Reg. Soc. Sci. Upsaliensis 9, No. 11 (1935), and C. B. van den Berg, thesis, Gröningen University, 1957; M emission from Skinner, Bullen, and Johnston, Phil. Mag. 45, 1070 (1954); and M absorption from Tomboulian, Bedo, and Neupert, J. Phys. Chem. Solids, 3, 282 (1957).

[84] Conventionally, this position in the uncorrected absorption curve is taken as the energy of *the* absorption edge (although the first inflection point in the *transmission* curve is sometimes used, with even less theoretical justification).

[85] D. E. Bedo and D. H. Tomboulian, Bull. Am. Phys. Soc. Ser. II, 3, 192 (1958); see also Tech. Rept. "The $M_{2,3}$ valence band spectrum of manganese" (July, 1958), Contract DA-30-115-ORD-669.

[86] We are currently in the process of repeating the work of Hanson and Herrera[40] on the K spectra of copper and nickel. The self-absorption correction is not as large as they reported since (1) they did not first correct the observed emission for instrumental resolving power, (2) they used an experimental curve for the absorption edge before it had been corrected for the instrumental resolving power, and (3) their correction for

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

satellites is very much stronger than is apparent before the correction for self-absorption is made, and careful correction for the satellites becomes imperative.[87] We would expect that, after all the pertinent corrections have been made, any remaining overlap of the emission and absorption spectra (at the Fermi energy) should be accounted for as the effects of thermal motion of the conduction electrons and of lattice vibrations in the emitting anode and in the absorber.

We may remind ourselves that, by application of the selection rule to the principal jumping electron, the K emission refers to transitions to a final state having an electron vacancy of p-type symmetry in valence band. In this application, the wave-function symmetry of the valence vacancy is viewed from the position of the K shell of the emitting atom. The K-absorption process leaves the system with one inner K vacancy and with the ejected electron having p-type symmetry whether it is in an excitation orbital or in some unoccupied part of the conduction continuum. In the L_{III} and M_{III} emission or absorption spectra, the final state vacancy or the ejected electron, respectively, has either s or d symmetry.[88]

As a help in the interpretation of the experimental curves, we may first become acquainted with the theoretical prediction (one-electron model) of the density of states in the solids under study. Such a prediction, which neglects the perturbation of the inner vacancy, is sketched in Fig. 12. This is a "composite" of calculations by several investigators.[89] The double maximum in the $3d$ band is predicted by all the calculations. The base width of the theoretical $3d$ band seems to be about 2.7 ev for nickel (Koster) and about 3.4 ev for copper (Howarth). Extrapolation of these calculated results to zinc is probably not justified because of its different crystal structure. The Fermi energy is indicated for nickel and copper, and it is estimated for zinc. Ten electrons go into the bands for nickel, 11 for copper, and 12 for zinc. In Fig. 12, the Fermi energies are placed for the respective areas to be in this ratio. As drawn, the Fermi energy for copper is about 2 ev above the top of the $3d$ band and also about 2 ev below the bottom of the $4p$ band.

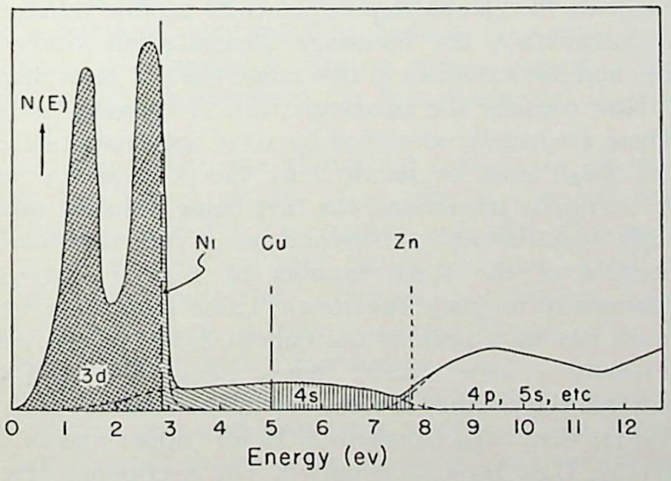


FIG. 12. "Theoretical" density of valence states for metallic nickel and copper, and (questionably) for zinc. This is a composite sketch from several different calculations based on the one-electron model.[89] Estimated Fermi energies are indicated.

Because of the dipole selection rule, we expect quite different shapes in the K emission on one hand and in the L_{III} and M_{III} emission on the other hand, neglecting excitation spectra in each case. Insofar as Fig. 12 applies and insofar as the observed spectra show the normal unperturbed states of the solid, the L_{III} and M_{III} curves should both reflect the s and d band structure of Fig. 12, weighted heavily in favor of d because of the relative transition probabilities and because of the greater density of states, whereas the K curves should show merely the admixed p symmetry in these bands. (In K-quadrupole radiation, the transition of s symmetry to d symmetry is allowed but, as mentioned in the foregoing, the intensity of this radiation is very faint relative to dipole radiation.)

First consider the L_{III} and M_{III} emission. The successive shift of the observed L_{III} and M_{III} emission from the Fermi energy, as we go from nickel to zinc, seems to confirm the $3d$-electronic band as most probably containing the final electron vacancy. On the other hand, it is most disconcerting that there is no (or very little) evidence in these observed curves for the theoretically predicted double maximum in the $3d$ band. The full widths at half-maximum of the observed L_{III}-emission bands are about 2, 3, and 1 ev, respectively, and predicted width at half-maximum of the $3d$ band is about 2 ev for nickel (Koster) and a little wider for copper (Howarth). No theoretical prediction is available for zinc. Even after allowing about 0.5 ev (very rough) in the observed curves for resolving power, etc., the width agreement is fairly good, somewhat unexpected and probably fortuitous in view of the double-peak disparity. The widths at the base are not in agreement. Moreover, the observed M widths at half-maximum are uniformly greater than 2 ev. There is no obvious reason in the one-electron metal why the L_{III}-emission shapes should be different from the M_{III} shapes. And the very wide K bands are not easily explained, even after an estimated 3 ev (very rough)

the background radiation (the side of the nearby intense β_1 line and the continuous radiation) was not quite right.

[87] The serious tangle of satellites as often reported in spectral curves recorded photographically [e.g., see Y. Cauchois, Phil. Mag. 44, 173 (1953), and references cited therein] is frequently a consequence of overexposure of the photographic film and a failure to correct properly to a linear intensity scale. But the correction for satellites is usually not an easy one because of the poorly known background intensity, a background that may contain essential structure itself.

[88] Again note that, at least in hydrogenic theory, the transition probability for p to s is much less than for p to d.

[89] E.g., see E. Rudberg and J. C. Slater, Phys. Rev. **50**, 150 (1936); J. C. Slater, Phys. Rev. **49**, 537 (1936); G. F. Koster, Phys. Rev. **98**, 901 (1955); G. C. Fletcher, Proc. Phys. Soc. (London) **A65**, 192 (1952); D. J. Howarth, Proc. Roy. Soc. (London) **A220**, 513 (1953); C. Kittel, *Introduction to Solid State Physics* (John Wiley and Sons, Inc., New York, 1956), Chap. 12.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

allowance is made for the correction for resolving power, self-absorption and satellites.

Considerable point is usually made of the apparent coincidence in the energy positions of the abrupt high-energy side of emission and the low-energy side of absorption for the metals sodium, magnesium, and aluminum. This coincidence is always sought for good metals, but it is not so readily seen for most metals. The curves for nickel, copper, and zinc are typical.

In general, we find some spotty agreements in terms of the theoretical predictions based on the one-electron model; but more disagreements. Perhaps some of the apparent difficulties would disappear if proper corrections were made to the observed curves, when more sanitary conditions and materials are used, etc.[90]

In particular, the correction for excitation spectra may be by no means negligible. For example, it is likely that the excitation orbital which forms inside the occupied valence band (see earlier discussion of the $K\beta'$ line for elements in this region of atomic numbers) is not similarly occupied in all K states in the emitting anode of the x-ray tube. If this is the case, then the K emission may be broadened as a result of the energy separations of overlapping excitation states, and also may be accompanied by the $K\beta''$ "satellites."[38,44]

Similar arguments apply to the L_{III} and M_{III} emissions for which the energy separations of the respective excitation states are expected to be somewhat smaller than for the K states. Further, the relative probability of the jump of a normal valence electron, compared with the jump of an excitation valence electron, may possibly be greater in the L_{III} than in the M_{III} emission because of the overlap of the respective wave functions and because the perturbation is stronger in the L_{III} case. If this is so, and if the final-state vacancy following the excitation electron jump is always at a particular favored place in the valence band (as mentioned earlier as a possibility), then we might expect the L_{III} emission to be narrower than the M_{III} emission and to be displaced slightly in energy position. We may continue this speculation: If the favored position of the positive hole in the final state is indeed responsible for the L_{III} emission's being narrower than the M_{III}, and also responsible for the relative energy positions of the L_{III} and M_{III} bands, then perhaps we may conclude that, after the excitation electron has jumped into the L_{III} vacancy, the positive hole is near the top of the $3d$ band. However, in the case of nickel, the L_{III} emission is only a little narrower than the M_{III}, and there is no significant difference in the peak positions. Such speculations are interesting, but very little significance should be attached to their conclusions at this time.

Also contributing to a difference in the energy positions and shapes of the emission bands is the difference in lifetimes of the respective initial states. Thus, in these states, effects of phonon interactions and degrees of lattice relaxation are different.

The L_{III} emission for nickel and copper shows in each case a small bump on the high-energy side of the Fermi energy. It is likely that this bump is a conventional satellite. It is possible, but perhaps not likely, that much of the asymmetry in the L_{III}- and M_{III}-emission bands for copper and zinc is due to high-energy satellites. Satellites, being essentially atomic in character, may generally be identified by observations with elements in a mono-step sequence of atomic numbers. Unfortunately, the necessary identification study of $L\alpha_1$ and $M\gamma$ satellites in this range has not been made.

Now consider the amazingly wide K-emission bands. These are usually identified by x-ray spectroscopists as the $K\beta_{2,5}$ lines or bands, *viz.*, the $K \rightarrow M_{IV,V}$ and $K \rightarrow N_{II,III}$ transitions, the first being nominal quadruple radiation and the second being very improbable because of the small number of $N_{II,III}$ electrons (nominally no such electrons). These bands are very faint, having a peak of only about 1/100 of the peak intensity of the nearby $K\beta_{1,3}$ lines, $K \rightarrow M_{II,III}$, whereas the number of electrons in the nominal $M_{IV,V}$, $N_{II,III}$ (i.e., $3d_{3/5,5/2}$) shells is 10 for copper and in the $M_{II,III}$ (i.e., $3p_{1/2,3/2}$) is only 6. The K-emission bands are obviously severely suppressed but, even so, are more intense than would be expected for normal quadrupole radiation. We conclude that the emission is actually due to an admixture of p-type symmetry in the nominal $3d$, $4s$ energy bands, but that this admixture is not very great or else the K bands would be more intense than they are. The relative feebleness of this $K\beta_{2,5}$ radiation is believed to be an important point in regards the degree of admixture in general.

Mott[91] has suggested that, from an interpretation of the van der Waals forces, an appreciable percentage of

[90] The curves reproduced in Fig. 11 are believed to be the most reliable reported to date for these spectral regions. However, the following is illustrative of the experimental situation. The K-emission curves of Bearden and Shaw[83] have been challenged by Hanson and Herrera[40]; the L_{III}-emission band of copper has been reported by several investigators with large disagreements (compare the two curves by Cauchois [Y. Cauchois, Phil. Mag. 44, 173 (1953), and references cited therein] and by Shaw *et al.*[83] as shown in Fig. 24 of reference 3 with the curve in Fig. 8 of van den Berg[83]); and the $M_{II,III}$ emission is in similar trouble (compare Fig. 3 of Skinner *et al.*[83] with Fig. 3 of E. M. Gyorgy and G. G. Harvey [Phys. Rev. 87, 861 (1952); see also Phys. Rev. 93, 365 (1954)] and with Fig. 2 of D. E. Bedo and D. H. Tomboulian [Phys. Rev. 113, 464 (1959)], and see Bedo and Tomboulian[85] who suggest that the assumed background is often highly arbitrary. And the L_{III}-absorption curve of our Fig. 11 has unfortunately a rather vaguely specified ordinate scale.

It is possible that the largest single factor in the experimental disagreements is the unknown intensity scale in much of the photographic work. For example, Cauchois reports that the measured L_{III} band width (and presumably the shape and the apparent structure) for nickel depends upon the exposure time. This dependence is probably not due to a change in materials or structure of the emitting anode of the x-ray tube; rather, the general situation seems to be identical to that in the early work in measuring the relative intensity of satellites {e.g., compare Fig. 1 of F. K. Richtmyer and R. D. Richtmyer [Phys. Rev. 34, 574 (1929)] with Fig. 3(b) of reference[62]} in which discrepancies of a hundred percent were not uncommon.

[91] N. F. Mott, Phil. Mag. 44, 187 (1953).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

admixed p symmetry may exist at and just below the Fermi energy in both nickel and copper. This is not believed to be likely because of the low intensity of the K emission, but if it were it would possibly account for the extra bump of intensity observed at -2 ev for copper, and it would presumably give rise to an abrupt edge in the nickel emission somewhat as is observed for, say, aluminum. Such an edge is not found for nickel.

The observed K emission above the Fermi energy may perhaps be relatively too intense to be accounted for as satellites.[92] It is barely possible that there is an appreciable number of stray electrons hovering around in the crystalline anode just above the Fermi surface; if so, and if enough of these hovering electrons drop into the K vacancies, we may be able to explain such a high-energy tail. It is also barely possible that such transitions take place in the case of the L_{III} and/or M_{III} emission. Such transitions, however, would seem to be very improbable.[93]

The large intensity maximum in the K-emission band near the Fermi energy in zinc asks for explanation. As we have stated above, it is likely that the density of states and the Fermi energy as displayed in Fig. 12 are not applicable to zinc. In particular, it may be that the low-energy tail of the $4p$-electronic band is occupied in zinc, and that the observed intensity maximum in question reflects this occupation as dipole radiation. If this is so, and if excitation states do not invalidate the one-electron model, the zinc K emission should exhibit the abrupt high-energy edge expected for metals under these conditions; perhaps it would if the corrections were properly made to the curve. But, even so, it is probable that an energy gap of a few volts would exist between the K emission and absorption edges.

Turning now to the absorption curves, the "jump-ratio" for nickel at the M_{III} edge is about 2 but decreases to about 1.3 for copper and zinc. This may be taken as evidence that the $3d$-electronic band in nickel is not quite fully occupied, but that in both copper and zinc it is fully occupied. Figure 12 is, of course, in accord with this evidence.

Obvious component structure appears in the absorption curves near and above the position of the Fermi energy. It is not clear whether this structure corresponds simply to a variation in the unperturbed $N_a(E)T(E)$ curve or is evidence of excitation states, plasma oscillations and multiple-inner-vacancy states.[94]

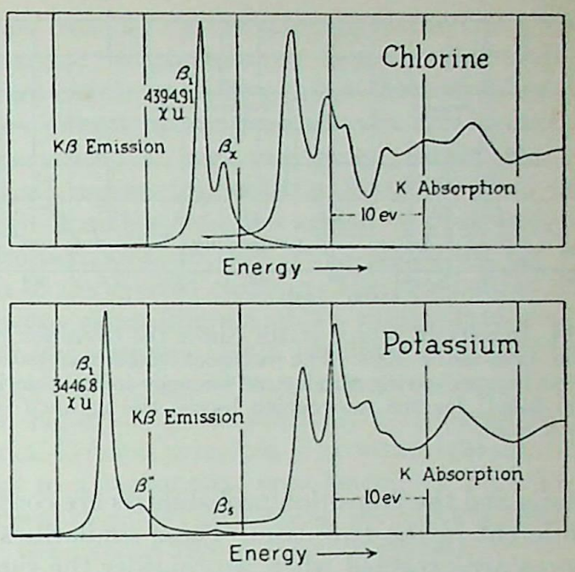


FIG. 13. K emission and absorption spectra (uncorrected) for crystalline potassium chloride.

The band structure of the transition metals is of considerable current interest, primarily because of their electrical and magnetic properties. In this interest, correlations of the results of x-ray measurements have been barely attempted[95] but the possibilities are very promising, especially after the observed curves for selected solid materials have been properly corrected for the various effects (a) through (h) as listed earlier.

Potassium Chloride

Figure 13 shows uncorrected curves for an insulator.[96] Here, the valence band in crystalline potassium chloride, nominally the Cl $3p$ electrons, is probably so far removed from the Fermi energy that the K emission band, designated β_1, is very narrow; the observed full width at half-maximum of β_1 is 1.21 ev. On the high-energy side of β_1, there is observed the line β_x for chlorine and β'' for potassium, and also a rather long faint tail in each case. β_x or β'', and the tail, are probably a complex of satellites and, if so, are of no great interest to us now.

In absorption, the excitation states are no doubt responsible for at least the most prominent maxima—indeed, the overlapping of the excitation spectra and the unperturbed solid-state band spectra is so great that we are at a loss to say where any feature of the unperturbed solid-state spectra may be.

The absorption structure is roughly similar for either the chlorine or the potassium inner vacancy although

[92] Stephenson [S. T. Stephenson, Phys. Rev. 58, 877 (1940)] reports that much of this apparent satellite structure may be an inverse view of the Kronig structure brought out in the self-absorption process. If so, this structure should disappear when proper correction for self-absorption is made.

[93] Many investigators have speculated about this possibility in attempts to explain the apparent satellite structure: e.g., see Y. Cauchois, Compt. rend. 201, 1359 (1935), T. Hayasi, Sci. Rept. Tôhoku Univ. 25, 785 (1936), 31, 1 (1942), and van den Berg.[83]

[94] Photographically recorded CuK absorption curves[55] show in the range 0 to 15 ev four wiggles (component structure) not indicated in Fig. 11. We are currently restudying this region with a two-crystal spectrometer.

[95] E.g., see N. F. Mott and K. W. H. Stevens, Phil. Mag. 2, 1364 (1957).

[96] These curves are from the paper by Parratt and Jossem.[31] The method used in reference 31 for correcting the emission line widths for the instrumental resolving power is not valid at these wavelengths, as has been shown by the better method of reference 17. An interesting emphasis on the need of a *proper* correction in the absorption curves is demonstrated by Parratt, Hempstead, and Jossem [Phys. Rev. 105, 1228 (1957)] in the so-called "thickness effect."

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

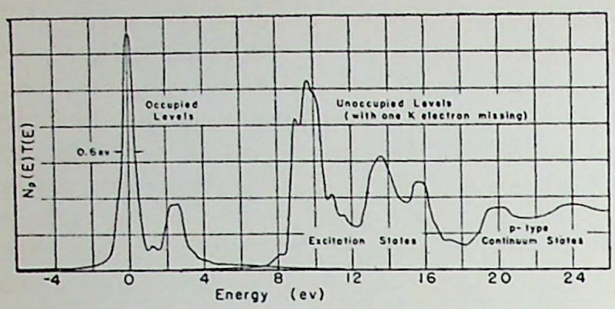


FIG. 14. Density of p-type states (times the transition probability) in crystalline KCl. The occupied electronic levels are those for a crystal having an electron vacancy in the chlorine $3p$ (valence) band; for the unoccupied levels, the vacancy is in a chlorine K shell.

the details and the transition probabilities are considerably different in the two cases. These similarities and differences are expected when we consider the similarities and differences in the immediate neighborhoods of the respective atoms in the crystal.

The chlorine curves for the other alkali chlorides also show[97] different details in the absorption structure, but all of these curves are characterized by pronounced maxima and minima. For LiCl and NaCl, the deep minimum that appears between the first two maxima for KCl and RbCl is effectively filled up, indicating either a smaller energy interval between the most prominent excitation states or else a greater number of such states. The structure for CsCl is again different, but the crystal structure of this chloride is body-centered-cubic whereas the other chlorides are of face-centered-cubic structure. The K-absorption curves for potassium in the potassium halides likewise show both similarities and differences.[98] Similar strong absorption maxima and minima for chlorine are also observed when the chlorine is in a gaseous phase, either in HCl or in free Cl_2[99] (see Fig. 3); this might be expected since the bound-ejected-electron excitation states are vestigial molecular or atomic states.

Figure 14 presents the $N_p(E)T(E)$ graph for potassium chloride. For this graph the chlorine K data shown in Fig. 13 have been fully corrected except for corrections (e), (g) and (h). The corrections for both kinds of resolving power were made by the method of Porteus and Parratt[17]: First, the observed transmission curve was corrected for the instrumental spectral window, and then the absorption curve (the logarithm of the thus-corrected transmission curve) was corrected for a Lorentzian K state.

Now, we see that the chlorine $3p$-valence band is indeed much narrower than has been customarily believed from early theoretical calculations (This problem is discussed in reference 31.) A recent calculation by Howland,[100] however, concludes that the base-width of this band is about 1.5 ev, in sensible agreement with the curve in Fig. 14. But, more recently, Howland has extended his theoretical calculations to predict a shape of this band as well as its width[101]; he predicts a half-maximum width of 1.32 ev, a value much greater than the observed 0.6 ev width, and he predicts a low-energy shoulder in the band which is much more prominent than the slight shoulder seen in Fig. 14.

This feature, just referred to as a slight shoulder, may be evidence of the spin-doublet states $3p_{\frac{1}{2}}$ and $3p_{\frac{3}{2}}$. The separation between these two states in gaseous atomic chlorine is about 0.11 ev and is expected to be about the same in crystalline KCl. It is possible to have two components of identical shape (near Gaussian) and separated by 0.11 ev, each with a width of about 0.5 ev, and with the $3p_{\frac{3}{2}}$ component on the high-energy side being twice as intense as the $3p_{\frac{1}{2}}$ component, such that their sum agrees with the curve of Fig. 14 within the accuracy with which the curve is known. Another interpretation of the slight shoulder would assign it to one or more excitation states of the valence-electron-configuration type, similar to the situation with the transition metals of the manganese group.

From the absorption part of the spectrum, it is significant that the number of evident excitation states is larger than has been previously recognized or theoretically discussed for x-rays. The prominent states here are no doubt of the bound-ejected-electron type.

The question arises as to the meaning of the width of a single component excitation state as it appears in the "unoccupied" part of Fig. 14. This width is the residual width after correction of the true spectrum for the "atomic" K state which is less wide than any K-excitation state. If a full-width K-excitation state had been used in the correction procedure, the result would have been degenerate (a Dirac delta-function[102]). The saving difference between these two K state widths is to be found primarily in the effects of phonons in the lattice. This needs further discussion.

Lattice Spacing and Phonons in KCl

First, we point out an effect due to the average value of the lattice spacing. Then we discuss the effect of statistical time and/or spatial variations in this spacing. The first is relatively easy to discuss insofar as we have knowledge; the second, the effect of the phonons, is a little more involved.

The energy of any x-ray state, i.e., of a system containing an inner electron vacancy, depends upon the total electric field in which the vacancy finds itself. As the lattice expands, e.g., by an increase in temperature, the contribution of electric field due to the presence of neighboring atoms (or ions) is generally altered, and so

[97] Parratt, Porteus, and Jossem (to be published).
[98] S. Kiyono, Sci. Repts. Tôhoku Univ. First Ser. 36, 1 (1952).
[99] Stephenson, Krogstad, and Nelson, Phys. Rev. 84, 806 (1951).
[100] L. P. Howland, Phys. Rev. 109, 1927 (1958).
[101] L. P. Howland, Quarterly Progr. Rept. Solid-State and Molecular Theory Group, MIT (January 15, 1958), p. 50.
[102] A Dirac delta-function $\delta(x)$ is zero for $x \neq 0$, infinite for $x=0$, and $\int \delta(x)dx=1$.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

is the energy of the x-ray state. The magnitude of the energy shift is probably small but is not easily deduced theoretically. It is believed that the shift is generally to a higher energy for x-ray states in atoms in a metal, and probably also for the anion in an ionic crystal, but to a lower energy for the cation in the ionic crystal.[103]

However, the ground state, i.e., the initial state in any absorption transition, is by definition not altered by the expansion of the lattice. Hence, the energy shift appears directly in the position of the main absorption edge. For a given solid it is expected that the shift is about the same for all states of energy greater than a few hundred ev; it is too difficult to predict the relative shift for states of less energy, e.g., for the valence state. Hence, for an emission line from a transition between states each of energy greater than a few hundred ev, the energy shift caused by the thermal expansion of the lattice, is expected to be small.

In absorption, a lattice spacing effect, in addition to the shift in energy of the main edge, comes into the Kronig and Hayasi structure since the energy positions in this structure depend upon Bragg reflection of the ejected electron in the lattice. The interference conditions in Bragg reflection are temperature sensitive since they depend upon the lattice spacing. The energy dependence of the interference conditions turns out to be an inverse quadratic function of temperature.

Incidentally, the argument may be advanced that a sensitive method for studying the chemical bonds around an atom in a solid, since these bonds affect the lattice spacing, may be in interpretations of the energy positions of (inner-state) $\rightarrow$ (valence-state) emission (e.g., $K \rightarrow V$), and/or of the inner-state absorption edge, as the chemical composition of the solid is varied. However, in compounds or alloys the role of the excitation states of both the valence-electron-configuration and the bound-ejected-electron types depend upon the chemical composition. It is likely that in a given alloy, even one having a uniform phase, the so-called normal valence configuration may not be the same for all atoms. Alterations in the excitation states probably have a small effect on the peak positions of emission lines but have possibly a large effect on the absorption spectra. In any case in such a chemical-bond study, they must be considered in addition to the effect due to the change in lattice spacing alone.[104,105]

Now consider the broadening of the energy of x-ray states due to phonons or lattice vibrations. (Some authors object to calling this a thermal effect because lattice vibrations are present even at absolute zero of temperature.) It was mentioned earlier that only a localized region of the solid is effectively involved in the energy of each state. This region is, of course, as great as the pertinent electron orbits, including the effective orbit of the ejected electron. The localization is due to the very short lifetime of an x-ray state and to the relatively long time for a phonon (or any other likely energy-bearing agent[106]) to travel through the lattice. This factor in localization is a consequence of the Franck-Condon principle. The x-ray system, therefore, refers to a region that may be only a very few lattice spacings in extent (excluding the ejected electron if it is in the remote part of the continuum).

The energy of a state is affected by the relative positions of all the electrons and nuclei in this localized region, and lattice vibrations give rise to fluctuations in this energy. Of course, the observed spectra represent a statistical average of the spectral energies from all the compressed or expanded regions that are emitting or absorbing the x-rays; the effect generally is to broaden an emission line, an absorption line, or an absorption edge. The statistical vibration broadening in either emission or absorption, considering all vibration frequencies and localized regions, is essentially Gaussian in shape.

For an ionic crystal, there are two kinds of lattice vibrations: (1) the so-called optical branch in which the adjacent positive and negative ions vibrate out of phase,[107] and (2) the acoustic branch in which adjacent ions vibrate approximately in phase. Vibrations frequencies extend from zero to about 10^{13} cps. Hence, after a time greater than about 10^{-12} or 10^{-13} sec, a completely new time and spatial distribution of compressed and expanded regions is expected.

For a metal crystal, the energy smear in an inner-vacancy state due to the vibrations in the electric field may be about a tenth of an ev or so at room temperature, and for an ionic crystal it may be somewhat greater because of the optical phonons. The smear is conveniently expressed as the root mean square of the

---

[103] Interesting studies of the temperature effects in K emission and absorption for ferromagnetic alloys have been reported by V. A. Kazantsev [Bull. Acad. Sci. USSR, Phys. Ser. **20**, 97 (1956), English translation, and Doklady AN SSSR **101**, 477 (1955)]. Although he gives simple interpretations in terms of the one-electron-jump type of diagram, the role of excitation states probably does not invalidate the general conclusions in regards the main trends of energy change.

[104] Nordling, Sokolowski, and Siegbahn report [Arkiv Fysik **13**, 483 (1958)] that the apparent K and L_{III} states for copper change in energy by 4.4 and 2.5 ev, respectively, from metallic Cu to CuO.

[105] Das Gupta reports [Phys. Rev. **80**, 281 (1950)] that the energy shift of the $K \rightarrow V$ or $L_{III} \rightarrow V$ emission peak correlates qualitatively with the heat of formation of a compound or alloy. This is in agreement with our expectations. However, in seeking quantitative agreement, Das Gupta assumes that the heat of formation is due entirely to the energy shift of the valence band alone, there being no shift in the inner state. This assumption does not seem reasonable.

[106] The electron vacancy in the x-ray state has some mobility in the lattice; this feature also extends the local region of interest. Not much is known [T. Takeuchi, Progr. Theoret. Phys. **18**, 421 (1957)] about the transfer of *inner* electron vacancies from one atom to the next, although this mobility is believed to be very small.

[107] The transverse mode of this type of vibration is excitable by an electromagnetic field, such as a light wave; hence the term "optical." We are concerned primarily with the longitudinal modes. The strongest polarization in the crystal occurs when adjacent ions vibrate 180° out of phase.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

amplitude of the vibrations of the state energy, and this for a Gaussian shape of the state is a little less than the half-width at half-maximum. For metallic potassium, the smear so expressed is calculated to be 0.15 ev at room temperature; for potassium chloride, the smear is about 0.15 ev at 0°K, about 0.16 ev at 100°K, about 0.19 ev at the Debye temperature, 230°K, and perhaps 0.23 ev at 373°K.[108] Such fluctuations are present in the energy of each x-ray ionization state. The magnitude and phase of the energy fluctuations are about the same for all ionization states in a given atom in the solid for which the effective diameter of the vacancy orbital is small compared with the atomic volume, i.e., for all states whose radiative lifetimes are less than about 10^{-13} sec.

It follows, then, as a prediction, that any emission line arising from a transition between ordinary x-ray states both of which have radiative lifetimes short compared with about 10^{-13} sec has an energy width that is rather insensitive to the lattice vibrations. But each absorption line, band or edge is expected to be broadened, regardless of the lifetime of the final state. Radiative lifetimes less than about 10^{-13} sec are expected for all x-ray states except those for atoms of the lowest atomic numbers and except for those states corresponding to a vacancy in the valence band.

For a singlet state of very short lifetime, shorter than about 10^{-13} sec, the shape is expected to be Lorentzian, as determined by the statistics of random-decay phenomena; and the shape of spectral lines (either emission or absorption) involving such states is also expected to be Lorentzian, and the shape of the *main* absorption edge, if the continuum of "unoccupied" states starts abruptly and is uniformly dense, is expected to be an arctangent. Indeed, within experimental error, the Lorentzian shape is observed for emission lines between *singlet* states both having very short lifetimes. However, in emission involving a longer lived state, e.g., the valence state, there is evidence that the tail on either side of the observed line is less high than Lorentzian, indicating a significant Gaussian contribution in the final state. (For example, the valence state derived from the β_1 line in Fig. 14 shows this effect.) In the longer wavelength range, Bedo and Tomboulian report[78] that the shape of the lithium K state, as interpreted from the K emission, is more nearly Gaussian than Lorentzian; this is to be expected if the K-state lifetime is of the order of, or greater than, about 10^{-13} sec,[109] a lifetime that is reasonable from other arguments.

Now consider a short-lived excitation state of the bound-ejected-electron type. The sensitivity of such a state to lattice vibrations depends upon the diameter of the orbit of the bound-ejected electron. If this electron remained very close to the position of the electron vacancy, its negative charge would effectively cancel the positive charge of the vacancy, and the energy of the state would suffer only minor fluctuations due to vibrations of the electric field. On the other hand, if the excitation orbit were large compared with the lattice spacing, such that the electric fields near a large number of atoms were sampled in the orbit, the lattice vibration sensitivity would be about the same as for an ordinary unexcited x-ray ionization state.[110] It is believed that most excitation orbits are of an intermediate size for which no prediction of the lattice vibration sensitivity is made at this time.

Another broadening effect in each excitation state, of either the valence-electron configuration or the bound-ejected-electron type, is due to the finite lifetime of the electron (or electrons) excitation orbital (or orbitals) as a consequence of all modes of decay other than the filling of the inner vacancy. The most likely types of such decay transitions are those due to Auger or phonon interactions. This type of broadening, Lorentzian in shape, is very small in the present KCl spectra, but the broadening by phonon interactions is possibly important when the inner vacancy has a longer lifetime, e.g., in absorption spectra in the long wavelength region.[111] However, even in the long wavelength

---

[108] These energy fluctuations are calculated from formulas supplied by Professor A. W. Overhauser. For the longitudinal modes in a monovalent metal,

$$[\langle (ev)^2\rangle]^{\frac{1}{2}} \doteq \left(\frac{NC^2k\theta}{6\rho u^2}\right)^{\frac{1}{2}}\left[1+\left(\frac{8T}{3\theta}\right)^2\right]^{1/8},$$

where N = number of atoms/cm³, ρ = density in g/cm³, u = velocity of longitudinal sound in cm/sec, T = absolute temperature, θ = Debye temperature, k = Boltzmann constant, and C = Fermi energy; and for the longitudinal optical modes in an ionic crystal of two types of ions,

$$[\langle (ev)^2\rangle]^{\frac{1}{2}} \doteq 8\frac{(3\pi^2)^{1/6}e^2}{n^2a^{7/2}}\left[\left(\frac{M_2}{M_1}\right)^{\frac{1}{2}}+\left(\frac{M_1}{M_2}\right)^{\frac{1}{2}}\right]\left(\frac{h}{\pi\rho\omega}\right)^{\frac{1}{2}}\left(\frac{2}{\exp(h\omega/\pi kT)-1}+1\right)^{\frac{1}{2}},$$

where e = electronic charge in esu, n = optical index of refraction, a = lattice constant, M_1, M_2 are masses of the two kinds of ions, h = Planck constant, ω = average optical mode frequency, and the other quantities are as defined in the foregoing. The particular type of mode for which the equation is devised is probably the most important one in the respective type of crystal. Inclusion of the other modes, e.g., transverse modes, may increase the total smear somewhat but probably not by more than the uncertainty in the value calculated from the one equation. This uncertainty is estimated to be less than a factor of 2, and the indicated trend with temperature is believed to be very much better than a factor of 2.

[109] Or, if the valence state, i.e., the *final* state in this Li K emission, has a long lifetime, the emission may also be predominantly Gaussian as a consequence of this lifetime and lattice vibrations, independent of the shape of the K state. If the proposal of Skinner-Landsberg[32] for metals is correct, the lifetime of the low-lying valence states for lithium are shortened by Auger transitions and the shape of such a low-lying state in the band becomes essentially Lorentzian.

[110] This presumes that the dynamic phonon-electron interaction is negligible during the short lifetimes of interest; but the statistical static interactions cause a broadening which may increase or decrease the effect in the ordinary unexcited x-ray ionization state.

[111] This broadening has been calculated by A. Radkowsky [Phys. Rev. 73, 749 (1948)] and found to be appreciable for

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by eGangotri Foundation USA

region, this broadening effect may be overpowered by the vibration Gaussian contribution which is also a consequence of the longer state lifetime.

Excitation States for KCl in Absorption

Our discussion of the excitation states shown in Fig. 14 may now be completed. The "atomic" K state used in the correction procedure was obtained, as discussed in connection with Table I, from $K\alpha_1$-emission lines and from the strong resonance absorption line in gaseous argon. Because of short lifetimes, this state does not contain the broadening effects of the phonons. So corrected, each component excitation state in Fig. 14 is expected to show phonon width and to be essentially Gaussian in shape, the width being roughly the same for all excitation states. The structure shown in Fig. 14 is consistent with this view. The band of excitation states about 4 ev wide in the center of the figure evidently contains many—at least six—overlapping component states, and the very steep sides of this band indicate component widths of the order of 0.2 ev. Such a width is in agreement with the calculated phonon width given above for KCl at room temperature. The rugged structure extends over a range of about 10 ev, and the total number of excitation states is probably greater than 15. Beyond about 20 ev from the valence band, the density of states has relatively small undulations in it, roughly the sort of undulations expected in the theoretical Kronig structure.

An experimental study of the chlorine K absorption in potassium chloride has been carried out at temperatures from 100°K to 373°K.[112] Two of these curves in Fig. 15, not corrected for resolving power, show: (1) a general shift of the order of 0.2 ev to lower energies, except for the first maximum in the region labeled p-type continuum, (2) broadening of the excitation states, and (3) some alterations in the component structure.[113] It seems reasonable to suppose from these

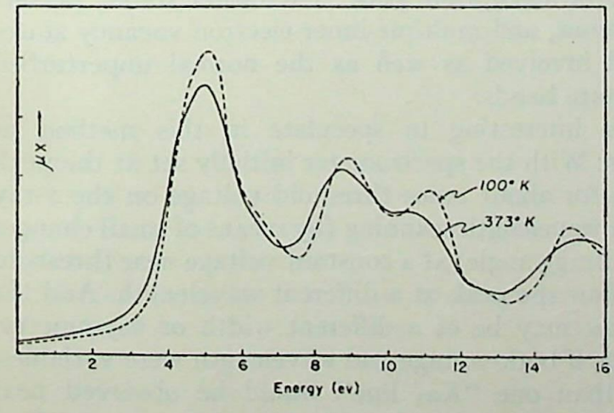


FIG. 15. Temperature dependence in the chlorine K-absorption spectrum of crystalline KCl.

results that the excitation electron in the first excitation state is essentially shared with the near-neighbor potassium ions, in agreement with the usual theoretical model.

A feature of great solid-state interest in a curve of the type shown in Fig. 14 is the magnitude of the energy gap between the normal valence band and the bottom of the conduction continuum. We discuss this feature now. Nominally, the bottom of the continuum refers to the states of s-type symmetry whereas Fig. 14 gives the density of states of p-type symmetry. Final states of s (or d) symmetry are involved in the ultraviolet absorption, and the desired energy gap has long been interpreted from this absorption as 9.46 ev in crystalline potassium chloride.[114] This interpretation, although the numerical value of 9.46 ev may possibly be changed to about 8.3 ev as a result of recent work,[115] would place the bottom of the s-type continuum at about the position of the central band of excitation states in Fig. 14. Admixture of p type in the nominal s-type states is no doubt present, and some fraction of the density of states in the region labeled excitation states in Fig. 14 should therefore be interpreted as part of the normal continuum. There is no obvious feature of the curve by which we can identify either the bottom of the continuum or even the existence of the continuum as some appropriate fraction under the excitation states. Possibly the deep minimum at about 16 ev above the valence states represents the bottom of the p-type continuum. States to the right of this minimum are designated p-type continuum states in the figure but we do not wish thereby to imply that the continuum necessarily begins here.

In general spectroscopy, it is common to add the energies of two or more selected lines to get the energy of a third; e.g., $(K \rightarrow L_{III})+(L_{III} \rightarrow M_{III})=(K \rightarrow M_{III})$. By a similar argument,[116] in which the energy positions of absorption edges are included, we may possibly derive some help in identifying the gap between the valence states and the bottom of the continuum. In KCl the valence states are the Cl M_{III} states, and we may write the gap-equations as

$$\begin{aligned}\text{gap (to } p\text{-type continuum)} &= K_{\text{abs}}-(K \rightarrow M_{III}) \text{ from Fig. 14} \\ &= M_{I\ \text{abs}}+(L_{III} \rightarrow M_I) \\ &\quad +(K \rightarrow L_{III})-(K \rightarrow M_{III}) \quad (1)\end{aligned}$$

[112] Parratt, Hempstead, and Porteus (to be published).

[113] The observed excitation states in ultraviolet absorption in KCl are also temperature sensitive as to both the apparent energy-positions and widths [Hartman, Nelson, and Siegfried, Phys. Rev. 105, 123 (1957)]. The vibrations in the electric field in the region of a valence orbital are believed to be too complex for even qualitative prediction at this time.

[114] E.g., see N. F. Mott and R. W. Gurney, *Electronic Processes in Ionic Crystals* (Oxford University Press, London, 1948), second edition, p. 95.

[115] The arguments of E. A. Taft and H. R. Philipp [J. Phys. Chem. Solids 3, 1 (1957)] may be applied to the experimental curves of Hartman, Nelson, and Siegfried [Phys. Rev. 105, 123 (1957)] for potassium chloride.

[116] The author is indebted to Professor E. L. Jossem for developing this argument in the present application.

The argument is essentially the Ritz-combination principle mentioned earlier as being valid in a gross description but not valied in detail. Our present interest is gross.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

gap (to s-type continuum)

$$= M_{III\ \mathrm{abs}} \text{ from ultraviolet absorption}$$
$$= L_{III\ \mathrm{abs}} + (K \rightarrow L_{III}) - (K \rightarrow M_{III}). \quad (2)$$

The only real help to be gained in this type of argument lies in a less ambiguous recognition of the appropriate[117] M_I and L_{III} absorption energies than is the case for K absorption in Fig. 14—the excitation spectra are expected to be somewhat different because of the different positions of the electron vacancy. The M_I state has the same symmetry as the K state, *viz.*, s type, but the L_{III} state has p-type symmetry. The gap-equations (1) and (2) are written separately for the p type and for the s-type continua because it is believed that the experimental values may well be different for these two gaps, and it is alleged theoretically that the admixture of p-type symmetry at the very bottom of the nominal s-type continuum is most unlikely.

The energies of all three chlorine emission lines in the gap-equations have been accurately measured for KCl —$K \rightarrow M_{III} = 2815.12$ ev, $L_{III} \rightarrow M_I = 182.75$ ev, and $K \rightarrow L_{III} = 2622.23$ ev—but the chlorine L- and M-absorption edges for KCl have not been recorded. The Cl $M_{I\ \mathrm{abs}}$ for LiCl has been determined by O'Bryan[118] who remarks that the absorption bands "show a width from 2 to 4.5 ev at the head with less intense continuous absorption extending to shorter wavelengths." The energy at the center of the strong absorption band is 29.7 ev. O'Bryan's description of the absorption curve also fits fairly well the Cl K-absorption curve for LiCl as reported by Parratt *et al.*[97] It is reasonable to conclude that 29 ev is probably a minimum value for the Cl M_I absorption energy for KCl. The Cl L_{III}-absorption energy is also to be deduced indirectly from measurements, and, again, only an approximate value can be obtained. We use the Cl L_{III}-absorption energy for gaseous CCl_4, 200.6 ev, as reported by Prins,[119] and add to this energy about 3.5 ev for the following reasons: 1.5 ev are added as a guess as to the separation between the first strong absorption peak and the appropriate absorption edge. The remaining 2 ev comes from a comparison of the Cl K-absorption spectra for gaseous CCl_4 by Nilsson[120] with the Cl K absorption for KCl by Parratt *et al.*[97]; the K edge shifts by about 2 ev in going from the gaseous CCl_4 to solid KCl. We assume that the L_{III} edge shifts by the same amount. Thus, the approximate energy of the Cl L_{III} absorption is taken to be 204 ev.

Substitution of these numerical values in Eqs. (1) and (2) gives

gap to p-type continuum = 19 ev
gap to s-type continuum = 11 ev.

The uncertainty in the 11-ev value may be 2 or 3 ev, and the 8 or 9 ev previously reported for the s-type case may possibly be a better value than the 11 ev. The 19 ev for the p-type case was deduced as a *minimum* value, and it also may be uncertain by a few ev. The M_I- and L_{III}-absorption edge energies are based on uncorrected curves recorded photographically and, in some instances, with low instrumental resolving power. Nevertheless, the conclusion that there is a difference of several ev between the effective bottoms of the s- and p-type continua is experimentally the best we have to date. This conclusion may easily be altered to say that there is only a very small amount of p admixture in the difference region.

These arguments are interesting and suggest desirable supplementary measurements for the future. At the present time, they are not sufficiently quantitative to provide a convincing basis for the energy-gap interpretation of the curve of Fig. 14; they do however tend to confirm the division of the curve into the general regions labeled excitation states and p-type continuum.

Other Types of X-Ray Curves

There are other ways in x-ray spectroscopy to study the electronic band structure of solids, but these we barely mention.

One way is to record and interpret a curve of the emission intensity of a characteristic line, say the $K\alpha_1$ line, with each of various values of the anode voltage in the neighborhood of the threshold excitation voltage.[121] This method purports to examine the unoccupied levels of the solid by inquiring as to just where and with what probability the inner electron can be ejected to them by cathode-electron impact. It is analogous to the x-ray absorption method except that the radiative selection rules do not apply. When scrutinized more carefully, the correlation between electron energy and the emitted x-ray intensity is as complex as the interpretations in absorption work: Excitation states, plasma oscillations, and multiple-inner-electron vacancy states are all involved as well as the normal unperturbed solid-state bands.

It is interesting to speculate in this method as follows: With the spectrometer initially set at the peak of $K\alpha_1$ for about twice threshold voltage on the x-ray tube, a wavelength scanning (by means of small changes in the Bragg angle) at a constant voltage near threshold may show the peak at a different wavelength. And the line now may be of a different width or asymmetry. Perhaps if both voltage and wavelength were variables, more than one "$K\alpha_1$ line" would be observed near threshold voltage; if so, they would correspond to different radiative combinations of K and L_{III} normal and excitation states. It is believed likely, however, that only one broadened $K\alpha_1$ line would be observed in this fashion because, if more than one were to appear,

[117] The difference between the "unexcited" x-ray level (the level referred to earlier with the subscript τ) and the energy of the absorption edge is presumed in this argument to be roughly the same for the various cases.
[118] H. M. O'Bryan, Phys. Rev. **49**, 879, 944 (1936).
[119] J. A. Prins, Physica **1**, 1174 (1934).
[120] A. Nilsson, Arkiv Mat. Astron. Fys. **29A**, 4, No. 28 (1943).
[121] A. Nilsson, Arkiv Fysik **6**, 513 (1953).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

then, in an appreciable fraction of the emission transitions, (1) the bound-ejected electron would have to jump differently by a few ev or (2) the valence-electron configuration would have to change differently. With regard to the latter possibility, it may be noted that the time required in the formation of the perturbation due to the inner vacancy near threshold voltage is somewhat longer than it is with much excess voltage on the x-ray tube, and consequently the respective probabilities for producing the various excitation states, especially of the valence-electron-configuration type, are perhaps altered.

Another method of studying the electronic band structure of solids is by the spectral details observed near the short-wavelength limit of the continuous x-radiation.[122] The cathode electron comes to rest, or nearly so, wherever it can find an acceptable unoccupied level in the solid. The x-ray intensity pattern gives information concerning the positions of the respective levels and the density of states. In this method, the excitation states are probably not involved, but the plasma oscillations and the multiple-vacancy states are present to complicate the interpretations.

[122] S. T. Stephenson, "The continuous x-ray spectrum," in *Handbuch der Physik*, S. Flügge and Marburg, editors (Springer-Verlag, Berlin, 1957), Vol. 30, pp. 368–370, and references cited therein.

VI. FURTHER READING

For further discussion of the experimental techniques, types and details of observed curves, variety of solids that have been studied to date, theoretical calculations, different speculations in interpretations, etc., the reader is perforce referred to the literature. Only a very few examples have been selected for inclusion in this review. Fortunately, several summaries of the general subject have recently appeared, each giving extensive references.[1–6]

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

Properties of Rutile (Titanium Dioxide)*

F. A. GRANT

National Bureau of Standards, Washington, D. C.

(Received December 5, 1958)

CONTENTS

1. Introduction 646
2. Crystal Structure 647
2.1. Introduction 647
2.2. Crystal Structure 647
2.3. Unit Cell Parameters 648
2.4. Bond Lengths and Angles 648
2.5. Interatomic Distances from Ionic Radii 648
2.6. Other X-Ray Investigations 648
2.7. Madelung Constant 648
2.8. Neutron Diffraction 649
3. X-Ray Properties 649
3.1. Soft X-Ray Spectra 649
3.2. X-Ray Absorption 650
4. Criteria of Bond Type 650
4.1. Introduction 650
4.2. Dielectric Constant 650
4.3. Electronegativity 650
4.4. Magnetic Susceptibility 650
4.5. Bond Lengths 650
4.6. Fourier Projection 650
4.7. Solubility 650
4.8. Pauling's Rules 651
5. Parameters for the Evaluation of Experimental Data of Semiconductors 651
5.1. Introduction 651
5.2. Donor Sites in Rutile 651
5.2.1. Noninteracting Donor Sites 652
5.2.2. Donor Bands 652
5.3. Interpretation of Hall Data 652
5.3.1. Noninteracting Donor Sites 652
5.3.2. Donor Bands 652
5.3.3. Hall Mobility 652
5.4. Interpretation of Electrical Conductivity Data 653
5.4.1. Intrinsic Semiconductors 653
5.4.2. Impurity Semiconductors 653
5.5. Interpretation of Thermopower Data 653
5.5.1. Impurity Semiconductors 653
5.5.2. Correlation of Thermopower and Conductivity Data 654
5.5.3. Correlation of Thermopower and Hall Data 654
6. Electrical Conductivity 654
6.1. Rutile Single Crystals (Intrinsic) 654
6.1.1. Field Strength Dependence 655
6.1.2. Current *versus* Time 655
6.1.3. Ionic Conductivity 655
6.1.4. Dependence on Area of Cross Section 655
6.1.5. Other Phenomena 655
6.2. Rutile Ceramic (Intrinsic) 655
6.3. Reduced Rutile 656
6.3.1. Experimental Results 656
6.3.2. Single Crystals *versus* Ceramics 656
6.4. Rutile Containing Added Impurities 657
6.4.1. Conductivity in Air at Normal Pressure 657
6.4.2. Impurities of the Type X_2O 657
6.4.3. Impurities of the Type XO 657
6.4.4. Impurities of the Type X_2O_3 658
6.4.5. Impurities of the Type XO_2 658
6.4.6. Impurities of the Types X_2O_5 and XO_3 658
6.4.7. The Results of Hauffe and Co-Workers 658
6.4.8. Effect of Impurity Concentration 659
7. Thermopower 659
7.1. Introduction 659
7.2. Experimental Results 659
8. The Hall Effect 660
8.1. Hall Data of Reduced Rutile 660
8.2. Hall Mobility 661
8.3. Magnetoresistance 662
9. Magnetic Susceptibility 662
9.1. Introduction 662
9.2.(a) Lattice Susceptibility 662
9.3.(b) Carrier Susceptibility 663
9.4.(c) Impurity Contribution 663
9.4.1. Electrons in Impurity Bands 663
9.4.2. Donor Sites in Nonstoichiometric Rutile 664
9.4.3. Rutile Containing Impurities 665
9.5.(d) Surface Contribution 666
10. Thermal Conductivity, Specific Heat, and Various Physical Properties 666
10.1. Thermal Conductivity 666
10.2. Thermal Coefficient of Expansion 666
10.3. Specific Heat and Debye Temperature 666
10.4. Miscellaneous Physical Properties 667
10.4.1. Density 667
10.4.2. Melting Point 667
10.4.3. Compressibility 667
10.4.4. Hardness 667
10.5. Miscellaneous 667
11. Dielectric and Optical Properties 667
11.1. Dielectric Constant and Loss Angle 667
11.2. Static Dielectric Constant and Polarizability 668
11.3. Optical Absorption 669
11.4. Color 670
11.5. Refractive Indices 670
11.6. Reflection Coefficient 670
12. Adsorption and Surface Layers 670
12.1. Photoconductivity in Surface Layers 671
13. Band Structure of Rutile—Conclusions 671
14. Thermodynamic Properties and Reduction 672
14.1. Thermodynamic Properties 672
14.2. The Reduction of Rutile 672
14.2.1. Heating in Oxygen at Reduced Pressure 673
14.2.2. Use of the CO-CO_2 Equilibrium 673
14.2.3. Use of the Hydrogen-Water Vapor Equilibrium 673
14.2.4. Use of the Alkaline Earth Metals 673
14.2.5. Other Methods 673

1. INTRODUCTION

OXIDE semiconductors, such as rutile, differ from the elemental semiconductors, silicon and germanium, in several respects. Probably the most important is the high melting point, which increases the difficulty of obtaining specimens of high purity. A related effect is the high concentration of vacancies which is "frozen" into the specimens at elevated temperatures. Oxygen vacancies are believed to exceed the cation vacancies in rutile. The formation of an oxygen vacancy in the otherwise pure oxide provides

* Work supported by the U. S. Atomic Energy Commission.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

a source of two electrons having an activation energy less than the energy required to raise an electron to the conduction band in the stoichiometric material. Thus the transport properties of the oxide are altered by departures from stoichiometry, just as they are by the presence of impurities.

The second respect in which oxide semiconductors differ from silicon and germanium is the separation of the cations by the oxide ions. This separation leads to a reduction of the overlapping of the cation wave functions. It is undecided whether the wave functions of the cation 3*d* electrons are to be considered as overlapping, or "noninteracting." This question has been discussed by Verwey, *et al.*,[1] Mott,[2] Zener and Heikes,[3] and very recently by Morin.[4] It seems probable that the titanium 3*d* levels (or band) do not overlap the bottom of the much wider 4*s* conduction band. To the extent that electrical conduction is due to 3*d* electrons, one expects a low mobility compared with the electron mobility in silicon and germanium.

Finally, one has the difficulty of distinguishing between chemisorbed oxygen and oxygen from the adjacent bulk material. In other words, our concept of a surface is less precise than in the case of elemental solids.

For these reasons, it is appropriate and timely to examine critically all the available literature on the physical properties of rutile. In the following pages these properties are interpreted in terms of two electronic models. The attempted correlation of the various types of measurement (electrical conductivity, thermopower, etc.) by the use of these models is not satisfactory. It was therefore decided to discuss many physical properties that are not obviously significant from the point of view of the two models.

2. CRYSTAL STRUCTURE

2.1. Introduction

Titanium dioxide exists in three crystalline modifications, rutile, brookite and anatase, all of which have been prepared synthetically. Titanium dioxide precipitated from sulfate or chloride solutions at room temperature is amorphous even after drying at 110°C.[5] At elevated temperatures anatase is precipitated from sulfate solutions, but hydrolysis by direct boiling of chloride solutions produces rutile as the initial crystalline product. Other forms are converted to rutile when heated to temperatures between 700 and 920°C. The transition is not reversible. The stability of anatase is increased by the presence of 0.1% of certain anions.[6] The preparation and mutual transformation of these oxides are discussed in Barksdale.[7]

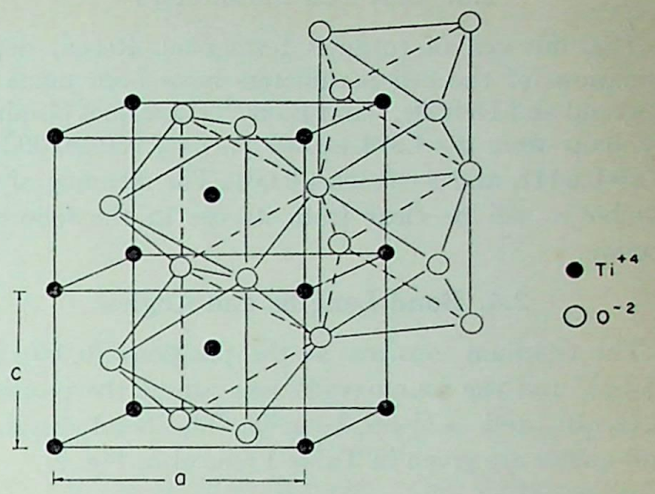


FIG. 1. Crystal structure of rutile (TiO_2). (Courtesy of R. G. Breckenridge, W. R. Hosler, and *The Physical Review*.)

Moore[8] has described the method of preparation of rutile single crystals by the Verneuil flame fusion method. When removed from the furnace the crystals are opaque black because of a slight deficiency of oxygen, but have the rutile crystal structure. After heating in a stream of oxygen the crystals become transparent, with a slight yellow coloration.

2.2. Crystal Structure

The crystal structure of rutile is usually described in terms of an ionic model based on the ions Ti^{4+} and O^{2-}. Although Syrkin and Dyatkina[9] consider that rutile may be regarded as intermediate between ionic and molecular in type, one notes that the Ti—O separation in the O—Ti—O grouping is actually slightly greater than the remaining four Ti—O spacings. This increased separation has been explained[10] as resulting from the repulsion between one oxygen ion and the four equidistant oxygen ions in the same octahedron.

The oxygen ions are arranged in the form of somewhat distorted octahedra (Fig. 1). Each octahedron shares one edge with adjacent members of the chain. Alternatively, the crystal structure may be visualized as chains of ions —O—Ti—O—O—Ti—O. All chains in the same layer are parallel, and the chains in adjacent layers are perpendicular to one another and to the *c* axis.

---

[1] Verwey, Haayman, and Romeyn, Chem. Weekblad 44, 705 (1948).

[2] N. F. Mott, Can. J. Phys. 34, 1356 (1956); Nuovo cimento Suppl., Ser. 10, 7, 312 (1958).

[3] C. Zener and R. R. Heikes, *Conference on Magnetism and Magnetic Materials*, Boston, Massachusetts, October 16-18, 1956 (Institute of Electrical Engineers, New York, 1957).

[4] F. J. Morin, Bell System Tech. J. 37, 1047 (1958).

[5] Seppo Wilska, Acta Chem. Scand. 8, 1796 (1954) (in English).

[6] H. Knoll and U. Kühnhold, Naturwissenschaften 44, 394 (1957).

[7] Jelks Barksdale, editor, *Titanium, its Occurrence, Chemistry and Technology* (Ronald Press, New York, 1951).

[8] C. H. Moore, Jr., Mining Trans. 184, 194 (1949).

[9] Y. K. Syrkin and M. E. Dyatkina, *The Structure of Molecules and the Chemical Bond* (Butterworths Scientific Publications, London, 1950).

[10] H. Remy, *Treatise on Inorganic Chemistry* (Elsevier Publishing Company, Amsterdam, 1956), Vol. II, p. 59.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

2.3. Unit Cell Parameters

The unit cell of rutile is tetragonal. Recent determinations of the cell parameters have been made by Legrand and Delville,[11] and Baur.[12] The values obtained by Baur were $a=4.594\pm0.003$ A, $c=2.959\pm0.002$ A, $c/a=0.6441$, and $x=0.306\pm0.001$. The meaning of the symbol x will be clear from its use in the following section.

2.4. Bond Lengths and Angles

The titanium ions are at the positions (0,0,0) and $(\frac{1}{2},\frac{1}{2},\frac{1}{2})$, and the four oxygen ions are at the positions $\pm(x,x,0)$, and $\pm(\frac{1}{2}+x,\frac{1}{2}-x,\frac{1}{2})$. The bond distances and angles are given in Table I (see also, Fig. 2).

2.5. Interatomic Distances from Ionic Radii

The ionic radii (Zachariasen quoted by Kittel) of the cation and anion are, respectively, $R_C=0.60$ A, and $R_A=1.46$ A.

In ionic crystals, the interatomic distance D_N is given by the relation $D_N=R_C+R_A+\Delta_N$, where N is the coordination number of the cation, and Δ_N is a correction which depends upon N. In rutile the cation has six nearest O^{2-} neighbors, hence $N=6$. From collected crystallographic data of other compounds, Δ_6 is found to be zero, or more correctly, R_C and R_A are calculated on the basis that $\Delta=0$ for $N=6$. The Ti—O separation calculated in this way (2.06 A) is somewhat greater than the observed values (1.944 and 1.988 A), indicating a slight departure from the rule. It is usually assumed that such a discrepancy reflects the contribution of the covalent forces to the crystal bonding.

TABLE I.

Bond	Type[a]	Number of bonds of its type per unit cell	
	(a)		
Ti—O	1	4	1.944±0.004 A
Ti—O	2	2	1.988±0.006 A
O—O	3	8	2.780±0.002 A
O—O	4	2	2.520±0.012 A
O—O	5	2	2.959±0.002 A

Angle	Bond[a] types	Number of angles of its type per unit cell	
	(b)		
O—Ti—O	1, 1 (smaller angle)	2	80.8°
O—Ti—O	1, 1 (greater angle)	2	99.2°
O—Ti—O	1, 2	8	90.0° (by symmetry)

[a] The numerals refer to the bonds indicated in Fig. 2.

[11] C. Legrand and J. Delville, Compt. rend. **233**, 944 (1953).
[12] W. H. Baur, Acta Cryst. **9**, 515 (1956).

2.6. Other X-Ray Investigations

By the use of Weissenberg photographs, Baur[12] has been able to carry out a Fourier projection of the elementary cell of rutile on the plane perpendicular to [001] (Fig. 3). This figure shows that the electron density does not fall to zero at any point between the oxygen and Ti_{II} ions. The "bridge" between cations and anions is an indication of a covalent contribution to the bonding. These results have been corroborated by the results obtained by a group of Spanish workers.[13,14]

2.7. Madelung Constant

Assuming purely ionic bonding and regular octahedra as the building blocks, Born and Bollnow[15,16] carried out a Madelung calculation for rutile. This analysis predicted a maximum Madelung constant when the axial ratio c/a was equal to 0.721. The observed ratio

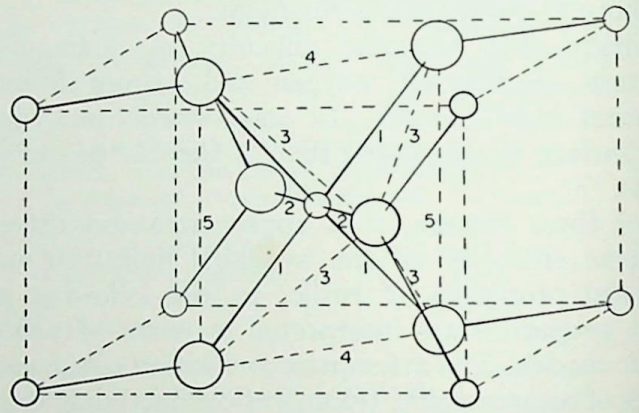


FIG. 2. The atomic arrangement of the tetragonal crystal rutile. Large circles represent oxygen atoms, small circles titanium atoms. The numbers refer to the bond types listed in Table I.

is $c/a=0.6441$. The calculated value of the Madelung constant is 4.816.

Making the assumption that the bonding is completely ionic (based on Ti^{+4} and O^{-2} ions), and taking into account also the repulsive contributions to the total energy, Lennard-Jones and Dent[17] have calculated the expected value of the parameter x (see "Bond Lengths and Angles"). In their method the same fixed separation r was assumed for all six Ti—O bonds, and the potential energy of the crystal was calculated as a function x. The value of the parameter x, necessary to minimize the potential energy was then obtained. Repeating this procedure for a number of values of r, a curve of x *versus* r was plotted. Using the experimental

[13] Bru Villaseca, Cubero, and Vega, Anales real soc. españ. fis. y quim. (Madrid) **46A**, 317 (1950).
[14] M. Cubero and Hernandez Montis, Anales real soc. españ. fis. y quim. (Madrid) **48A**, 133 (1952).
[15] M. Born and O. F. Bollnow, Naturwissenschaften **13**, 559 (1925).
[16] J. E. Lennard-Jones and B. M. Dent, Phil. Mag. **3**, 1204 (1927).
[17] J. E. Lennard-Jones and B. M. Dent, Phil. Mag. **3**, 1204 (1927).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

data available at that time, SnO_2 and PbO_2 were found to lie on the theoretical curve. Using Baur's[12] experimental value of $x=0.306\pm0.001$ for rutile, one obtains a value of $r=2.05\pm0.03$ A from the theoretical curve, which is somewhat higher than Baur's experimental value of 1.988 ± 0.006 A. Once again, one finds a discrepancy which probably reflects a covalent contribution to the bonding in rutile.

2.8. Neutron Diffraction

Using thermal neutrons, Shull and Wollan[18] have obtained Laue diffraction patterns of a number of single crystals, including rutile.

3. X-RAY PROPERTIES

3.1. Soft X-Ray Spectra

Confirmation of the band structure theory of solids is obtained from soft x-ray emission experiments. The probability of a transition from an electronic state in a higher band to a vacant lower state depends, among other factors, on the number of electrons in the upper state. Consequently the distribution of intensity with wavelength provides information concerning the density of states as a function of energy.

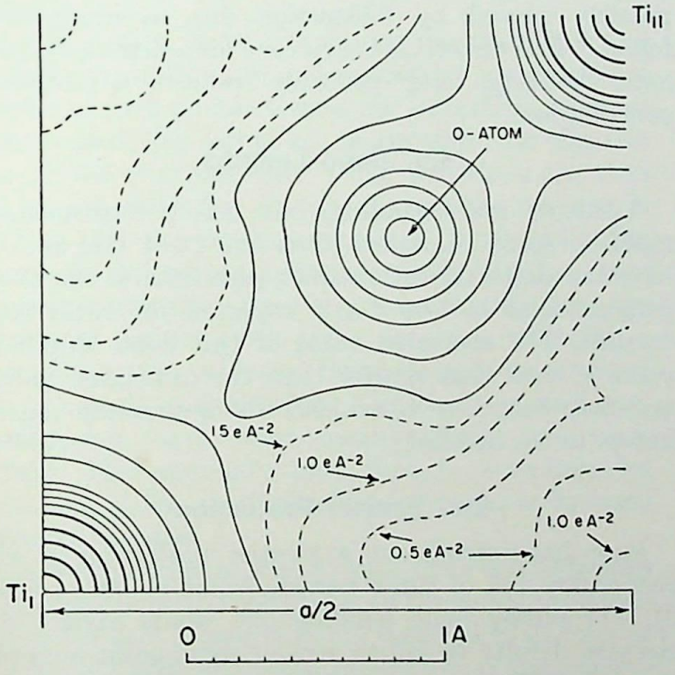


FIG. 3. Fourier projection of rutile on the [001] plane (after W. H. Baur[12]).

Soft x-ray absorption experiments also provide information concerning the energy levels in a solid. The K-absorption edge corresponds to the transition (consistent with the selection rules) of a $1s$ electron to the lowest empty level. Experimental values of the absorption coefficients are plotted against photon energy in Fig. 4.

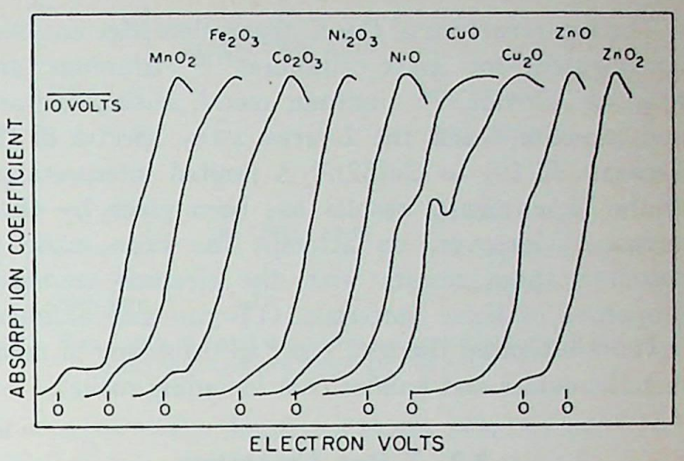


FIG. 4. K-edge structures for oxides of some of the transition metals (after Hanson and Knight[20]).

The distinguishing features of these curves are a small low-energy absorption, followed by a steep rise in absorption coefficient. On the high-energy side of the absorption edge the curve exhibits a fine structure characteristic of the environment of the atom (the type of bonding, etc.).

The shapes of such curves are determined by the absorbing atoms, their immediate surroundings, and by type of structure of which they form a part. Kiestra[19] has summarized his conclusions in the following words:

"(a) In a region close to the edge, the behavior of the absorption coefficient is mainly a property of the atom in question.

"(b) In the following region the fine structure is determined by the immediate surroundings of the atom.

"(c) In the region of highest energy, the fine structure depends essentially on the whole crystal lattice.

"For the transition metals and their compounds (a) comprises the region up to about 40 eV from the edge, (b) the region between about 40 eV and 150 eV, and (c) the region above 150 eV."

Recently, Hanson and Knight[20] have investigated the K-absorption edge structure of the fluorides, oxides, and sulfides of a number of transition metals. These workers found that oxides in which the metal assumed a higher valence exhibited greater low-energy absorption than the same metals in their lower valence form. In highly ionic compounds, the low-energy absorption is less than in compounds between elements exhibiting less difference in electronegativity.

When it is suspected that there is an appreciable change in symmetry and crystal bonding, as in barium titanate at the Curie temperature, the investigation of the K-absorption edge of titanium might prove to be of interest.

[18] C. G. Shull and E. O. Wollan, Naturwissenschaften 36, 291 (1949).

[19] S. Kiestra, "Conference on the Application of X-ray Spectroscopy to Solid State Problems," University of Wisconsin, Madison, October 23–25, 1950, Report NAVEXOS P-1033, p. 36.

[20] H. P. Hanson and J. R. Knight, Phys. Rev. 102, 632 (1956).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

The fine structure of the K-absorption edge has been investigated for pure titanium,[21,22] titanium and titanium dioxide,[23,24] titanium metal, rutile, anatase, and brookite,[25] and the L-series x-ray spectra of the elements K(19) to Ge(32).[26] A limited interpretation of the experimental results has been given by these authors. Moreover, no attempt has been made to correlate these results with the electron transport properties of these materials. O'Bryan and Skinner[27] have investigated the soft x-ray spectroscopy of many metallic oxides and halides (not including rutile).

3.2. X-Ray Absorption

Brewster[28] lists values of the mass absorption coefficient of rutile. At the wavelengths 0.01, 0.1, 1, and 2 A, the values of the mass absorption coefficient are, respectively, 0.05434, 0.1790, 38.71, and 250.4, respectively. The linear absorption coefficient is equal to the product of the mass absorption coefficient and the density.

4. CRITERIA OF BOND TYPE

4.1. Introduction

When the covalent contribution to the bonding is small, as in the alkali halides, a quantitative estimate of its effect may be made. When a quantitative estimate cannot be made, however, it is usual to make use of certain criteria of covalency, such as the type of crystal structure, atomic radii, bond lengths, position in the periodic table and electronegativity difference of the components, Pauling's rules, etc.

It should be emphasized that the following criteria of the extent of the ionic contribution to the bonding are not necessarily in close agreement. Their value is therefore more qualitative than quantitative in nature.

4.2. Dielectric Constant

The optical and static dielectric constants of covalent crystals are almost equal. In contrast, the static dielectric constant of highly ionic crystals is considerably greater than the optical dielectric constant. The values for rutile single crystals are $K=173$ and $K_{opt}=8.4$ in the c direction, indicative of a strong ionic character.

4.3. Electronegativity

Mulliken[29–31] defines the electronegativity of an atom as the mean of its ionization potential and its electron affinity, thereby making use of observable quantities. The concept of electronegativity is less precise when more than one valence electron is involved.

Hannay and Smyth[32] have derived a formula which enables the extent of the ionic character of a bond to be estimated from the electronegativities of the bonded partners. This approach yields a value of 43% ionic character for the Ti—O bond in rutile, which is about the same as that of HF and KI.

4.4. Magnetic Susceptibility

The tetravalent titanium and the divalent oxygen ion both have noble-gas configurations. The *temperature-dependent* paramagnetic susceptibility of stoichiometric rutile should therefore be zero, whether the bonding is ionic or covalent. Feeble (Van Vleck) *temperature-independent* paramagnetism, however, is observed, possibly caused by "distortion due to interatomic forces.[33] This distortion is evidence for a departure from ionic bonding, and probably reflects a covalent contribution.

4.5. Bond Lengths

Assuming ionic bonding, but taking into account repulsive forces, Lennard-Jones and Dent (see Sec. 2) have calculated the cation-anion separation in the plane perpendicular to the c axis expected for rutile-type crystals. The measured value of this bond length in rutile is somewhat shorter than the calculated value, and may be taken as evidence for a covalent contribution to the bonding.

4.6. Fourier Projection

Baur has carried out a Fourier projection of the elementary cell of rutile perpendicular to [001] (Fig. 3). For purely ionic bonding one would expect the electron density to fall to zero at some point between the ions.[34] The considerably higher electron density between the oxygen and Ti_{II} ions in Fig. 3 is evidence for a covalent contribution to the bonding.

4.7. Solubility

The low solubility of rutile in polar solvents such as water, is taken as evidence for a considerable covalent contribution to the Ti—O bonding.

---

[21] Seljakob, Krasnikov, and Stellezkii, Z. Physik **45**, 548 (1927).
[22] G. Okuno, Proc. Phys.-Math. Soc. Japan **18**, 306 (1936).
[23] M. A. Blokhin, Doklady Akad. Nauk. S.S.S.R. **95**, 965 (1954).
[24] V. H. Sanner, Z. Physik **112**, 430 (1939).
[25] I. B. Borovskii, Compt. rend. acad. sci. U.R.S.S. **26**, 764 (1940) (in English).
[26] F. Tyren, Arkiv. Mat. Astron. Fysik **25A**, No. 32 (1937).
[27] H. M. O'Bryan and H. W. B. Skinner, Proc. Roy. Soc. (London) **A176**, 229 (1940).
[28] Gordon F. Brewster, J. Am. Ceram. Soc. **35**, 194 (1952).
[29] R. S. Mulliken, J. Chem. Phys. **2**, 782 (1934); *ibid.* **3**, 573 (1935).
[30] C. A. Coulson, *Valence* (Clarendon Press, Oxford, 1951).
[31] A. F. Wells, *Structural Inorganic Chemistry* (Oxford University Press, New York, 1950).
[32] N. B. Hannay and C. P. Smyth, J. Am. Chem. Soc. **68**, 171 (1946). This formula also appears in references 30 and 31.
[33] L. F. Bates, *Modern Magnetism* (Cambridge University Press, New York, 1951), pp. 44 and 272.
[34] For example see the projection of the sodium chloride lattice given on page 315 of reference 9.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

4.8. Pauling's Rules

One may consider the reduction in stability in the rutile structure resulting from the sharing of one edge by adjacent octahedra (Pauling's third rule[35]) to indicate a departure from purely ionic bonding. This reduction in stability is probably compensated by a corresponding increase in the covalent contribution to the bonding, made possible by this structure.

5. PARAMETERS FOR THE EVALUATION OF EXPERIMENTAL DATA OF SEMICONDUCTORS

5.1. Introduction

It is desirable, whenever possible, to present experimental data in terms of parameters derived from a (more or less simple) model. In the present case, neither the accuracy nor the consistency of the available experimental data permits a choice to be made between the commonly used models. For this reason, the properties of rutile are discussed in terms of two simple models.

In this section, these two models are examined, and the experimentally observed quantities are expressed in terms of such parameters as the ionization potential of the donors, the forbidden energy gap, the effective mass of the electrons, etc. These equations are then used as an indication of the appropriate manner in which to plot the experimental data. In the following sections the experimental data pertaining to the electron transport properties of rutile are introduced, and the parameters derived from these data are given. In a later section, the values of these parameters obtained from magnetic susceptibility, electrical conductivity, thermopower, and other measurements, are summarized. This summary may then be examined for indications of the most satisfactory model to be used in discussing rutile.

It is assumed (as in Fig. 5) that there is a forbidden gap of several electron volts between the valence and conduction bands.[36] Donor sites are believed to lie one- or two-hundred millivolts below the bottom of the conduction band. There may be overlapping of donor wave functions leading to the formation of donor bands. In this case the effective mass is expected to be much higher in the donor band than in the conduction band.

The various symbols used in later sections are defined below:

$n_0 = 2(2\pi mkT_0/h^2)^{\frac{3}{2}}$ = the constant 2.52×10^{19} electrons/cc and is defined for $T_0 = 300°$K. This quantity is the so-called "effective number of states."

n_c = the concentration of electrons in the conduction band (per cm^3). Electrons in donor bands, which may incidentally contribute to the conduction, are excluded.

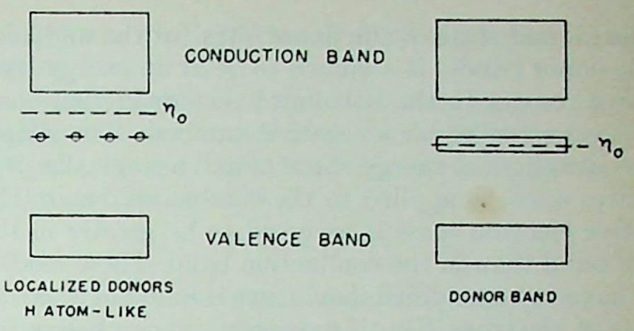


FIG. 5. Two types of band structure assumed in the discussion of the transport properties of rutile.

$M = (m_e/m)$ = the density-of-states effective electron mass.

$t = (T/300)$ = the absolute temperature divided by 300°K.

e = the electronic charge in coulombs.

η = the height of the Fermi level in joules, measured from the bottom of the conduction band.

k = the Boltzmann constant (units depend upon the context-joules per degree when transport properties are in question).

T = the absolute temperature in °K.

N_d = the concentration of donor atoms (per cm^3).

E_d = the energy of the donor sites with respect to the bottom of the conduction band, in joules.

b = the electron mobility in cm^2 volt^{-1} sec^{-1}.

E_g = the energy gap, i.e., the separation of the valence and conduction bands, in joules.

We take the criterion for the applicability of Maxwell-Boltzmann statistics to be $-\eta/kT > 3$.

5.2. Donor Sites in Rutile

Nonstoichiometry is believed to result in the presence of oxygen vacancies, together with twice that number of electrons, probably on Ti^{+3} sites.† No distinction will be made between nonstoichiometric rutile and rutile containing pentavalent impurities. In both cases the formation of donor sites will result. The presence of acceptor levels is considered if and when necessitated by experimental evidence. Two cases are considered corresponding to the two models: (a) noninteracting hydrogen-like donor sites, at sufficiently small concentration, in which case a single s electron is considered to move in the field of a net positive charge and (b) interacting donor sites, when the overlapping of the donor wave functions is appreciable. This interaction may lead to the formation of a donor band, and possibly to associations of hydrogen molecule-like donors, or associations of helium-atom-like donors and vacancies, in which two electrons are considered to move in the field of a net positive charge of two.

[35] R. C. Evans, *An Introduction to Crystal Chemistry* (Cambridge University Press, New York, 1952).

[36] W. Crawford Dunlap, Jr., *An Introduction to Semiconductors* (John Wiley & Sons, Inc., New York, 1957).

† The experiments of Gray and McCain indicate that heating of rutile in an atmosphere of hydrogen leads to a loss in weight of the rutile. This result is evidence for the creation of oxygen vacancies rather than hydrogen interstitial atoms. T. J. Gray, *The Defect Solid State* (Interscience Publishers, Inc., New York, 1957), p. 287.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

The ground state of the donor sites (or the midpoint of the donor bands) is assumed to lie at an energy level E_d with respect to the bottom of the conduction band (E_d is expected to be a negative number). The simple model of spherical energy surfaces and a single density-of-states mass, is applied to the conduction band. The effective electron mass is assumed to be greater in the donor band than in the conduction band. These models may have to be modified if evidence is found of a second source of electrons (i.e., if oxygen vacancies behave as helium-like centers).

Because of the random distribution of donors, the donor bands may have "tails" so that some overlapping of the conduction band may occur. However, this modification is not introduced into the mathematics because of the obvious difficulties. It can be shown[37] that these models lead to the following results:

$$n_c = n_0 M^{\frac{3}{2}} t^{\frac{3}{2}} \exp(\eta/kT). \tag{5-1}$$

5.2.1. Noninteracting Donor Sites

We usually make the additional assumption that the Fermi level η lies more than $3kT$ higher than the donor impurity levels.‡ This assumption implies that less than one impurity site in twenty is ionized, and leads to the following relation[38]:

$$n_c = \tfrac{1}{2} N_d \exp(E_d - \eta)/kT, \tag{5-2}$$

where N_d = the concentration of donor atoms (per cm^3), and E_d = the energy of the donor sites with respect to the bottom of the conduction band.

If Eqs. (5-1) and (5-2) are combined, one obtains the following expression for the concentration of conduction electrons in terms of the ionization potential of the donor impurity sites[39]

$$n_c = \left(\frac{n_0 N_d}{2}\right)^{\frac{1}{2}} M^{\frac{3}{4}} t^{\frac{3}{4}} \exp E_d/2kT. \tag{5-3}$$

5.2.2. Donor Bands

When the wave functions of the donor sites overlap, the energy levels of the isolated donors must be replaced by bands. Since there are N_d energy levels in the band, and each level may accommodate two of the N_d electrons, the donor band is only half filled at 0°K. Thermal excitations of electrons to the conduction band tend to lower the Fermi level as the temperature is raised. However, it can be shown that the Fermi level remains near the donor band provided that $N_d \gg n_c$. In this case it is appropriate to use the height of the Fermi level η as an indication of the position of the donor band. For this purpose η is expressed in terms of observable quantities by taking logarithms of both sides of Eq. (5-1) and multiplying by kT.

5.3. Interpretation of Hall Data

The Hall constant R for nondegenerate n-type semiconductors (discussed also in Sec. 8) is related to n_c by the relation[40]

$$n_c = r/eR = -6.25 \times 10^{18} r/R. \tag{5-4}$$

The magnitude of the coefficient r, which depends on the type of scattering,[40] lies between 1 and 2.

5.3.1. Noninteracting Donor Sites

Combining Eqs. (5-3) and (5-4) one obtains

$$1/(-R) = \text{constant} \times t^{\frac{3}{4}} \exp(E_d/2kT), \tag{5-5}$$

or

$$[\ln(-R) + \tfrac{3}{4} \ln t] = K_1 - E_d/2kT, \tag{5-6}$$

in which K_1 is a constant which is dependent on M, N_d and on the coefficient r of Eq. (5-4). When the quantity $[\ln(-R) + \frac{3}{4} \ln t]$ is plotted against reciprocal temperature, one obtains $-E_d$ from the slope, and if one knows N_d, and if M is temperature independent, one obtains M from the value of the intercept, K_1, made by this straight line.

5.3.2. Donor Bands

In Sec. (5.2.2) the height of the Fermi level has been discussed in the case in which the overlapping of the donor wave functions leads to the formation of donor bands. Provided that $N_d \gg n_c$, it may be shown that the Fermi level lies below, but near, the midpoint of the donor band. It is reasonable to assume that the electron mobility in the conduction band is much greater than in the donor band. It is therefore appropriate to plot the high temperature Hall data in terms of Eq. (5-1), which may be rewritten with the known quantities on the left-hand side, as follows

$$\ln n_c - \ln n_0 - \tfrac{3}{2} \ln t = \tfrac{3}{2} \ln M + \eta/kT. \tag{5-7}$$

At sufficiently low temperatures, however, as the concentration of conduction electrons decreases, the effect of the electrons in the donor band is expected to dominate. Maxwell-Boltzmann statistics are inapplicable to these latter electrons.

5.3.3. Hall Mobility

The electrical conductivity may be expressed

$$\sigma = n_c e b. \tag{5-8}$$

[37] See Eq. (5.33a) of reference 36.

‡ Because of the high value of the dielectric constant of rutile, E_d may prove to be so small that this assumption is justified only at very low temperatures, and for sufficiently high donor concentrations. The effect of the high dielectric constant will be counteracted to some extent if the effective mass ratio is also high. See p. 267 of reference 36.

[38] The justification for the factor $\frac{1}{2}$ in Eq. (5-2) found in Appendix III of E. Spenke, *Elektronische Halbleiter* (Springer-Verlag, Berlin, 1956).

[39] Apart from the factor 2, and the sign of E_d, this is Eq. (5-31) of reference 36. E_d, of course, is measured with respect to the bottom of the conduction band.

[40] Figure 6.53 of reference 36.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by eGangotri

Combining Eqs. (5-4) and (5-8) one obtains an expression for the mobility b in terms of the Hall coefficient and the electrical conductivity σ:

$$b=0.84\sigma(-R/r). \tag{5-9}$$

5.4. Interpretation of Electrical Conductivity Data

5.4.1. Intrinsic Semiconductors

If the density-of-states masses of electrons and holes are equal, then the Fermi level is in the middle of the energy gap, i.e., $\eta=-E_g/2$. Combining Eqs. (5-1) and (5-8) and

$$\sigma=n_0ebM^{\frac{3}{2}}t^{\frac{3}{2}}\exp(-E_g/2kT). \tag{5-10}$$

In this case the temperature dependence of the exponential factor is usually so great in comparison with that of the factors b and $t^{\frac{3}{2}}$ that Eq. (5-10) may be written

$$\sigma=K_2\exp(-E_g/2kT). \tag{5-11}$$

E_g may then be evaluated from the slope of a plot of $\ln\sigma$ *versus* reciprocal temperature.

5.4.2. Impurity Semiconductors

Isolated donors.—In the case of noninteracting impurities, one combines Eqs. (5-3) and (5-8) to give

$$\ln\sigma=\tfrac{1}{2}\ln(n_0N_d/2)+\ln e+\ln b+\tfrac{3}{4}\ln M+\tfrac{3}{4}\ln t+\exp(E_d/2kT). \tag{5-12}$$

Without some assumption concerning the temperature dependence of the mobility b, E_d cannot be evaluated from conductivity data alone. (The assumption of a single density-of-states electron mass for the conduction band necessarily implies that M is considered to be temperature independent.)

Breckenridge and Hosler[41] (following Fröhlich[42,43]) attempt to interpret their data by using the following relation for the mobility when the scattering is predominantly due to optical lattice vibrations:

$$b=A(\exp(\theta/T)-1), \tag{5-13}$$

where θ is the Debye temperature. After the appropriate substitution

$$\ln\sigma-\tfrac{3}{4}\ln t-\ln(\exp(\theta/T)-1)=K_3+(E_d/2kT), \tag{5-14}$$

in which K_3 is a constant which depends on N_d, A, and M. An assumption concerning the value of θ permits the left-hand side of Eq. (5-14) to be evaluated from conductivity measurements, and plotted against reciprocal temperature. From the slope E_d may be determined. Since K_3 is dependent on the unknowns N_d, A, and M, it is not feasible to attempt an interpretation of the intercept of this plot in the absence of further information.

Actually, (as shown in Sec. 8) the Hall mobility data of Breckenridge and Hosler for slightly reduced rutile are adequately represented above room temperature by the empirical relation

$$b=aT^{-2.5}, \tag{5-15}$$

where T is the absolute temperature, and a is a constant independent of temperature. Substitution of this relation into Eq. (5-12) enables one to write

$$\ln\sigma+1.75\ln t=K_4+E_d/2kT. \tag{5-16}$$

Donor bands.—At sufficiently low temperatures the electrons would be expected to lie in the donor band. The donor-band conductivity in this low-temperature range should depend on temperature chiefly through the mobility factor b. Little information would be obtainable concerning the position of the donor band with respect to the conduction band.

At higher temperatures the electrons in the conduction band should make a greater contribution to the conductivity than the electrons in the donor band. Provided that information is available concerning the electron mobility, the data on electrical conductivity at higher temperatures may be used to give information concerning the Fermi level, and hence the approximate position of the donor band. Making the usual assumptions concerning the applicability of Maxwell-Boltzmann statistics and of the concept of a single effective electron mass to the electrons in the conduction band, we may combine Eqs. (5-1) and (5-8) to yield the following relation:

$$\ln\sigma-\ln n_0e-\tfrac{3}{2}\ln t=\tfrac{3}{2}\ln M+\ln b+\eta/kT \tag{5-17}$$

in which the quantities on the left-hand side are all known. When the assumption implied in Eq. (5-13) is valid, we may combine Eqs. (5-13) and (5-17) and express the result in the form

$$\ln\sigma-\tfrac{3}{2}\ln t-\ln(\exp(\theta/T)-1)=K_5+\eta/kT. \tag{5-18}$$

When the $T^{-2.5}$ relation [Eq. (5-15)] is valid, one may write

$$\ln\sigma+\ln t=K_6+\eta/kT. \tag{5-19}$$

5.5. Interpretation of Thermopower Data

5.5.1. Impurity Semiconductors

Noninteracting donors.—Equating Eqs. (5-1) and (5-2) one obtains

$$n_0M^{\frac{3}{2}}t^{\frac{3}{2}}\exp(\eta/kT)=\tfrac{1}{2}N_d\exp(E_d-\eta)/kT \tag{5-20}$$

which may be expressed in the form

$$\tfrac{1}{2}\ln n_0+\tfrac{3}{4}\ln M+\tfrac{3}{4}\ln t+\exp(\eta/kT)=\tfrac{1}{2}\ln N_d+E_d/2kT-\ln 2. \tag{5-21}$$

Noting that

$$\eta/kT=Q/k+Q^*/kT \tag{5-22}$$

[41] R. G. Breckenridge and W. R. Hosler, Phys. Rev. **91**, 793 (1953).
[42] H. Fröhlich and N. F. Mott, Proc. Roy. Soc. (London) **A171**, 496 (1939).
[43] Fröhlich, Pelzer, and Zienau, Phil. Mag. **41**, 221 (1950).

CC-0. Bhabha Atomic Research Centre (BARC). Digitized by S3 Foundation USA

[see Eq. (7-2)], where Q^* is approximately equal to $2kT$, one obtains

$$eQ/k+\tfrac{3}{4}\ln(T/300)=K_7+E_d/2kT \qquad (5\text{-}23)$$

in which the constant K_7 includes Q^*/kT, and the impurity concentration term, and the effective mass term, which are assumed to be temperature independent. The thermopower, Q, is defined in Sec. 7.1.

Donor bands.—At sufficiently low temperatures, the contribution of the electrons in the conduction band is negligible. In this temperature region, the electrons in the donor band are expected to exhibit degeneracy, and the dynamic term Q^*/kT of Eq. (5-22) should make the major contribution to the thermopower. This contribution (approximately $2k/e$) is almost temperature independent.

At higher temperatures, where the electrons in the conduction band make the main contribution to the thermopower, the height of the Fermi level is related to the thermomower through Eq. (5-22). Replacing Q^* by its approximate value $2kT$, one obtains

$$E_f=\eta/e=(Q+2k/e)T. \qquad (5\text{-}24)$$

Thus when the donor sites interact to form donor bands one expects an almost temperature-independent value for QT in the higher temperature range. At still higher temperatures where thermal excitation raises a large proportion of the electrons from the donor band to the conduction band, the Fermi level falls below the donor band with increasing temperature, and Q should tend to a constant value.

5.5.2. Correlation of Thermopower and Conductivity Data

Impurity semiconductors.—When the temperature is sufficiently high for the contribution of the electrons in the conduction band to dominate the contribution to the conductivity of the electrons in the donor band, the correlation between thermopower and conductivity data is independent of the interaction between donor sites. Equations (5-1) and (5-8) may be combined to yield

$$\ln\sigma=\ln(n_0e)+\ln b+\tfrac{3}{2}\ln M+\tfrac{3}{2}\ln t+\eta/kT. \quad (5\text{-}25)$$

When the foregoing assumption, Eq. (5-13), concerning b is permissible, one obtains

$$\ln\sigma-\tfrac{3}{2}\ln t-\ln(\exp(\theta/T)-1)=K_8+\eta/kT \quad (5\text{-}26)$$

in which K_8 is a constant which depends on A and M. The thermopower Q may be expressed in the form $Q=(\eta-Q^*)/eT$, where Q^* is a quantity of the order of magnitude of $2kT$.§ Whence

$$\ln\sigma-\tfrac{3}{2}\ln t-\ln(\exp(\theta/T)-1)=K_8+(eQ/K)+Q^*/kT. \quad (5\text{-}27)$$

If an assumption is made concerning θ, the left-hand side of this equation may be evaluated from conductivity measurements. However, the quantity K_8 involves both the effective mass ratio M and the mobility factor A. Hence, in the absence of further information, no correlation between electrical conductivity and thermopower is possible.‖ To the extent that the term Q^*/kT is temperature independent, however, it should be possible to correlate dQ/dT with the rate of change of the left-hand side with temperature.

A similar conclusion is reached when the mobility b is inversely proportional to a power of the absolute temperature, as in Eq. (5-15).

5.5.3. Correlation of Thermopower and Hall Data

Correlation of thermopower Q with the Hall coefficient would involve one fewer variable (b), and would therefore be preferable to the correlation with electrical conductivity.

6. ELECTRICAL CONDUCTIVITY

6.1. Rutile Single Crystals (Intrinsic)

Cronemeyer[44] has measured the intrinsic conductivity of rutile single crystals in the temperature range 300° to 1400°C, using specimens approximately 1×1×10 mm in dimensions. The four-terminal method of measurement was used (see his Fig. 3); hence, the measured conductivity should be independent of the electrode material. However, the conductivity was found to be field dependent when the field strength was greater than 10 v/cm for the intrinsic conductivity, and greater than 0.1 v/cm for reduced (presumably nonstoichiometric) samples.

Cronemeyer's results may be represented by the following equations, in which σ represents the electrical conductivity in (ohm cm)$^{-1}$, and T is the absolute temperature:

Direction perpendicular to the c axis

623° to 1123°K

$$\ln\sigma=7.92-17\,600/T \qquad (6\text{-}1)$$

1123° to 1673°K

$$\ln\sigma=11.10-21\,200/T. \qquad (6\text{-}2)$$

Direction parallel to the c axis

773° to 1223°K

$$\ln\sigma=8.43-17\,600/T \qquad (6\text{-}3)$$

1223° to 1673°K

$$\ln\sigma=11.30-21\,200/T. \qquad (6\text{-}4)$$

‖ However, in some materials, such as NiO, it is possible to equate the density of states with the number of cations. This simplification, which results from the localization of the cation wave functions, facilitates a correlation of thermopower and conductivity data without requiring a knowledge of the effective electron mass. See F. J. Morin, Bell System Tech. J. 37, 1047 (1958).

§ References are given in Sec. 7.

[44] D. C. Cronemeyer, Phys. Rev. 87, 876 (1952).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

The respective activation energies calculated from these data are 1.53, 1.84, 1.53, and 1.84 ev. For intrinsic samples, the energy gaps are assumed to be approximately twice these values. In deriving these results the temperature variation of the $(T/300)^{\frac{3}{2}}$ and mobility factors of Eq. (5-1) have been neglected.

Cronemeyer associates the 3.06-ev energy gap obtained in this manner with the optical absorption edge at 3.02 ev.

A more recent investigation of the conductivity of single crystals of rutile has been carried out by Y. L. Sandler.[45] Preliminary published results indicated a somewhat higher conductivity in the temperature range 250° to 700°C, and a somewhat smaller energy gap of 2.8 ev.

6.1.1. Field Strength Dependence

Below 900°C, Cronemeyer's measurement[44] of the intrinsic conductivity of rutile showed some field strength dependence for fields greater than 10 v/cm. Measurements made at 820°C showed an increase in conductivity of about fifty percent when the field was increased to about 20 v/cm. Nonlinearity apparently begins at an even lower field strength (0.1 v/cm) in reduced samples.

6.1.2. Current versus Time

Sandler[45] has observed the current *versus* time characteristics of rutile single crystals. Measurements were made at temperatures up to 720°C for various applied steady voltages. Only curves for 654°C are given in the report. Although it was not explicitly stated, the two-terminal method was probably used, with evaporated gold electrodes. The discussion seems to indicate that the specimens were in an atmosphere of pure oxygen. To ensure that the observed time-dependent effects were not due to surface conduction, a guard-ring electrode was used.

The curves were characterized by relaxation times that decreased as the applied voltage increased. For 1.06-v potential difference, the relaxation time was about two minutes. The author considered that the data might be explained in terms of space-charge polarization and field emission.

6.1.3. Ionic Conductivity

Large currents could be passed through the rutile crystals for hours without destructive effect. Hence Cronemeyer concluded that the conductivity of rutile was essentially electronic.

6.1.4. Dependence on Area of Cross Section

The conductivity was observed to increase by a factor of fifteen when the cross-sectional area was increased from 3×10^{-2} to 6×10^{-2} cm^2. Cronemeyer suggested the possibility that this was due to "carrier injection with predominant recombination at the surface."

6.1.5. Other Phenomena

Sandler[45] has investigated the effect on the conductivity of rutile crystals of the heating during the process of depositing the evaporated gold electrodes. Electrodes were formed at 100 and 300°C. The conductivity ceased to be an exponential function of reciprocal temperature below 270°C, and below 500°C, respectively.

In the presence of water vapor, the conductivity was considerably increased. By the use of a guard electrode, it was shown that the effect was not due to surface conduction, but was probably due to the effect of moisture on the crystal-electrode interface.

6.2. Rutile Ceramic (Intrinsic)

Dependence of electrical conductivity on oxygen pressure is probably associated with departures from stoichiometry.¶ The conductivity of rutile containing intentionally introduced oxygen vacancies is discussed in a later section. The workers referred to in this section made no deliberate attempt to reduce their specimens more than the degree of reduction incidental to the conditions of measurement.

Earle[46] measured the electrical conductivity of pressed disks of rutile as a function of the pressure of the surrounding oxygen. His results may be expressed in the form

$$\ln\sigma = A - B\ln p, \tag{6-5}$$

where σ is the conductivity in (ohm cm)$^{-1}$, p is the oxygen pressure in atmospheres, and A and B are temperature dependent parameters.

At pressures below 30 mm of mercury B was found to be $\frac{1}{2}$ at 626°C, and equal to $\frac{1}{3}$ in the temperature range 755° to 968°C. Above 30 mm of mercury, B was equal to $\frac{1}{4}$. At 50 mm of mercury and 1000°K the conductivity was 1.6×10^{-6} (ohm cm)$^{-1}$. The activation energy was about 1.7 ev, and appeared to be independent of the oxygen pressure. Evaporated gold electrodes were used.

Earle investigated the possibility of ionic conductivity, and believes that its contribution was negligible under the conditions of his experiment.

Hauffe[47] also investigated the dependence of the electrical conductivity of rutile on the pressure of the surrounding oxygen gas. His results may be expressed by the same formula, but with different values of A

[45] Laboratory for Insulation Research, Massachusetts Institute of Technology, Progress Rept. No. XVII (1955).

¶ The effect of oxygen pressure on surface conductivity is also important. The effects of oxygen and other gases on the surface conductivity of silicon and germanium are discussed in *Semiconductor Surface Physics*, R. H. Kingston, editor (University of Pennsylvania Press, 1956).

[46] Marshall D. Earle, Phys. Rev. **61**, 56 (1942).

[47] K. Hauffe, *Reaktionen in und an festen Stoffen* (Springer-Verlag, Berlin, 1955), p. 136.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

and B. At 800°, 900°, and 1000°C, the values of B were, respectively, 1/4.8, 1/4.4, and 1/5.3.

Combining measurements made at these three temperatures, and using the approximate value of $\frac{1}{5}$ for B, Hauffe's results may be represented roughly by the relation

$$\ln\sigma = -\tfrac{1}{5}\ln p - 22\,400/T \qquad (6\text{-}6)$$

in which p is expressed in atmospheres, and T is the absolute temperature. The coefficient of $1/T$ yields an activation energy of 1.93 ev.

Gorelik[48] has measured the electrical conductivity of "pure" sintered rutile, and rutile to which various amounts of the oxides of the alkaline earths had been added. His results for "pure" rutile may be expressed by the following formula in the temperature range 380° to 490°K:

$$\ln\sigma = 2.5 - 11\,700/T. \qquad (6\text{-}7)$$

The corresponding activation energy is 1.02 ev. Gorelik's investigation of the effect of a magnetic field on the electrical conductivity is discussed in another section.

6.3. Reduced Rutile

The term "reduced rutile" is used here to signify that an attempt has been made to introduce oxygen vacancies by heating the specimen in oxygen at reduced pressure, or in an atmosphere of hydrogen or some other reducing gas. It is implied that the specimens have been quenched so that the vacancy concentration is higher than that corresponding to the equilibrium at the temperature and pressure at which measurements are made.

When an oxygen vacancy is created, the two electrons associated with the ion must be retained in the crystal to preserve electrical neutrality. There is some uncertainty whether these electrons (a) are associated with the oxygen vacancies to form helium-like donor sites, or (b) whether the electrons convert two Ti^{+4} sites into Ti^{+3} sites.

In case (a) one, or possibly two, activation energies are expected for electron transport phenomena when the vacancy concentration is too low for the formation of impurity bands. Because of the formation of donor bands one expects the disappearance of the activation energy at sufficiently low temperatures when the vacancy concentration is high.

6.3.1. Experimental Results

Breckenridge and Hosler[49] measured the electrical conductivity of reduced rutile single crystals and ceramics. The conductivity of these samples lay in the range 0.03 to 10 (ohm cm)$^{-1}$, depending on the degree of reduction, which was carried out in hydrogen at elevated temperatures.

[48] S. I. Gorelik, J. Exptl. Theoret. Phys. U.S.S.R. **21**, 826 (1951).

[49] R. G. Breckenridge and W. R. Hosler, Phys. Rev. **91**, 793 (1953).

The curves of the logarithm of the conductivity *versus* reciprocal temperature were fairly flat, and in most cases had negative slopes at 1000 and at 77°K.

There are so many disposable parameters in the simple model [Eq. (5-12)] (Debye temperature θ, impurity concentration, effective mass, etc.) that it is unlikely that its use in the interpretation of these data can yield significant information, except in conjunction with Hall data. The values of the electron mobility obtained from Breckenridge and Hosler's conductivity and Hall data are discussed in Sec. 8.

Kataoka and Suzuki[50] have made measurements of the electrical conductivity and thermopower on reduced rutile ceramic. When these results are plotted[51] according to Eq. (5-16) assuming isolated donor levels, the slopes correspond to values of the donor ionization potentials in the range 0.29 to 0.35 ev. An exception to this result (0.23 ev) was the sample of highest conductivity (20 ohm^{-1} cm^{-1}) for which the explanation could be partial degeneracy.

When the results are plotted[51] according to Eq. (5-19) assuming a donor band, the resulting values of the Fermi level lie in the range 0.12 to 0.15 ev, with the exception of the sample of highest conductivity which yielded a value of 0.09 ev.

Boltaks[52] and co-workers have investigated the conductivity of sintered specimens of rutile which had been reduced in hydrogen or in carbon monoxide. They found that the temperature variation of the pre-exponential term in Eq. (5-12) could not be neglected. When allowance was made for the $T^{\frac{3}{2}}$ factor in Eq. (5-12) a linear plot of $\ln\sigma$ *versus* reciprocal temperature was obtained. From the slope of the tangent to this curve, the ionization potential E_d of the donor sites was estimated to be less than 0.2 ev at room temperature. Apparently no consideration has been given to the effect of the variation of the mobility b with temperature [Eq. (5-12)].

6.3.2. Single Crystals versus Ceramics

The measurements of Breckenridge and Hosler[49] may be examined to compare the properties of polycrystal-

[50] S. Kataoka and T. Suzuki, Bull. Electrotech. Lab. Tokyo **18**, 732 (1954).

[51] Measurements of electrical conductivity made by Kataoka and Suzuki may be represented within the experimental error by the following expressions between $1000/T = 1.5$ and $1000/T = 3.5$:

Specimen	$\log_{10}\sigma$
1	$2.12 - 250/T$
2	$2.02 - 550/T$
3	$1.72 - 520/T$
4	$1.40 - 500/T$
5	$0.60 - 400/T$
6	$-0.13 - 475/T$

In the same temperature range, $\log t = \log(T/300)$ may be represented approximately as a function of $1/T$ by the expression $0.62 - 184/T$. Substitution of these expressions into Eqs. (5-16) and (5-19) yields the activation energies and ionization potentials given in the text. The mobility has been assumed to be proportional to $T^{-2.5}$ (see Sec. 8-2).

[52] Boltaks, Vasenin, and Salunina, Zhur. Tekh. Fiz. **21**, 532 (1951).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

TABLE II. Comparison of parameters for single crystals and ceramics reduced in hydrogen under approximately the same conditions (from Table I of Breckenridge and Hosler[a]).

	Reduction temperature °C	Reduction time in minutes	Resistivity at 300°K ohm-cm	Hall coefficient at 300°K $(-R)$	Activation energy ΔE_2 ev[b]	n_2[b]	Effective mass ratio of electrons m_e/m
National Lead single crystal	800	5	8	4	0.12	1.4×10^{19}	43
National Bureau of Standards ceramic	800	5	1	0.3	0.14	29×10^{19}	55
Linde single crystals (c direction)	800	indefinite	0.7	0.10	0.16	15×10^{20}	46
(a direction)	800	15	1	0.10	0.15	13×10^{20}	114
National Bureau of standards ceramic	800	average of 5, 10, 20 min specimens[c]	1	0.50	0.14	3.1×10^{20}	60

[a] See reference 49.
[b] See Eq. (8-2).
[c] The corresponding resistivity and Hall curves are almost coincident.

line specimens and single crystals. When this comparison is made for specimens that have been reduced under similar conditions, there is no evidence for a significant difference between the single crystal and polycrystalline specimens with respect to mobility, effective mass, or activation energy.

In Table II some of Breckenridge and Hosler's results have been tabulated for the purpose of comparing the parameters for single crystal and ceramics. The comparison is justified only for pairs of samples that had been reduced (in hydrogen) at the same temperature for approximately the same length of time.

As an example of the difficulty of interpreting these results, one notes that the Hall coefficient of the National Lead single crystal (reduced 5 min at 500°C) is higher (by a factor of eight) than that of the ceramic specimen subject to the same reduction. However, the Hall coefficient of the Linde crystal is less than or equal (depending on the direction of the current with respect to the optical axis) to that of the ceramic specimen. Similarly, there appears to be no evidence that the effective mass ratio is greater for single crystals than for ceramic specimens.

6.4. Rutile Containing Added Impurities

6.4.1. Conductivity in Air at Normal Pressure

The effect of impurities on the electrical conductivity of rutile ceramics has been investigated by Hauffe,[47,53] by Grunewald,[54] and by Johnson.[55]

Johnson's samples were air-quenched from 1200 to 250°C and measurements were made at the latter temperature. The results should be accepted with some caution because little attention was paid to the state of reduction of the specimens, and the impurities may have assumed valences other than those expressed in the formulas given. In every case the effect of the impurity approached a saturation value when the concentration exceeded about one or two mole percent. This saturation value is given in the following summary.

6.4.2. Impurities of the Type X_2O

The oxides of lithium, potassium, and silver decreased the conductivity of rutile, while that of sodium produced an increase (Table III). The effect was small in comparison with some of the results discussed in the following.

6.4.3. Impurities of the Type XO

The effects of admixture of the oxides of beryllium, magnesium, calcium, strontium, barium, zinc, cadmium,

TABLE III. Effect of various impurities[a] on electrical conductivity of rutile ceramic at 250°C (after Johnson[b]).

Impurity	Change in conductivity	Impurity	Change in conductivity
WO_3	×4000		
MoO_3	×0.5		
P_2O_5	×1020	BeO	×25
Sb_2O_5	×833	MgO	×0.5
V_2O_5	×27	CaO	×0.07 (?)[c]
Ta_2O_5	×4166	SrO	×0.25
Nb_2O_5	×5500	BaO	×0.95
		ZnO	×0.33
SiO_2	×3.3	CdO	×0.50
ZrO_2	×0.9	NiO	×0.33
B_2O_3	×6	CoO	×0.33
Al_2O_3	×0.4	Li_2O	×0.07
Fe_2O_3	×0.5	Na_2O	×1.7
Cr_2O_3	×3.5	K_2O	×0.7
Ga_2O_3	×0.4	Ag_2O	×0.7
Y_2O_3	×0.1		

[a] "Saturation" effect—usually about one mole percent impurity.
[b] See reference 55.
[c] See text.

[53] Hauffe, Grunewald, and Tränckler-Greese, Z. Elektrochem. **56**, 937 (1952).
[54] H. Grunewald, Ann. Physik **14**, 121 (1954); Hauffe, reference 47.
[55] G. H. Johnson, J. Am. Ceram. Soc. **36**, 97 (1953).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

nickel, and cobalt were investigated. The effect was small, except in the case of beryllium (increase by a factor of 25), and calcium (decrease by a factor of 0.07). The effect of beryllium was considered to be possibly due to the donor electrons resulting from the formation of interstitials. In the case of calcium the explanation of the anomalously low value may be as follows. Comparison of Johnson's Table IV with his Fig. 1 suggests that the correct value for calcium is probably considerably higher than 0.07, a value which was apparently influenced by one low experimental point.

Gorelik[56] has investigated the effect of various concentrations of the oxides of the alkaline earths on the electrical conductivity of rutile in the temperature range 380° to 490°K. Values of the activation energies obtained from plots of the logarithm of the conductivity *versus* reciprocal temperature lay in the range 0.78 to 1.15 ev for additions of CaO and SrO. At a temperature of 418°K the admixture of about three moles of CaO per hundred moles of TiO_2 caused a fourfold increase in conductivity. Further increase in CaO content caused the conductivity to decrease until it reached one quarter of its initial value when the CaO content reached twenty-five moles per hundred moles of TiO_2.

One mole of SrO per hundred of TiO_2 yielded a fifty-fold increase in conductivity over the value for "pure" rutile. Further admixture of SrO caused a decrease in conductivity, so that an admixture of 50 moles of SrO per 100 moles of TiO_2 resulted in a value of conductivity only slightly higher than the value for "pure" rutile. Gorelik interpreted his results in terms of contributions from ionic and electronic conductivity. This interpretation was based partly on his investigation of the effect of a magnetic field on the conductivity of these specimens, and is discussed further in another section.

6.4.4. *Impurities of the Type X_2O_3*

With the exception of the oxides of boron and yttrium, the effect of oxides of this group is not pronounced. The low value assigned to the oxide of yttrium appears to be unduly influenced by a single experimental point, and one could well expect a value as high as 0.3 from Johnson's Fig. 1.

Admixture of 0.5 mole percent of chromium oxide decreased the conductivity at 1100°C, but produced a sharp increase in conductivity at 500°C. Further increase in admixture of this oxide caused little change in conductivity. The apparent activation energy was reduced by a factor of three by the admixture of 0.5 mole percent. This apparent change in activation energy appears to be associated with a different mechanism of conductivity. This conclusion is strengthened by the oxygen pressure dependence of the conductivity of this material, discussed in another section.

[56] S. I. Gorelik, J. Exptl. Theoret. Phys. U.S.S.R. **21**, 826 (1951).

6.4.5. *Impurities of the Type XO_2*

The effect of the oxide of silicon appears to be surprisingly high in view of the valence assumed by this element.

6.4.6. *Impurities of the Types X_2O_5 and XO_3*

Some of the largest effects were produced by oxides in these groups. However, the low values corresponding to the oxides of molybdenum and vanadium are remarkable. The difference between the elements molybdenum (atomic number 42) and niobium (atomic number 41) is interesting, as is the difference between the effect of the oxides of tungsten (atomic number 74) and tantalum (atomic number 73).

Table IV summarizes the effect of adding certain oxides to rutile containing 0.5 mole percent of niobium oxide. The large effect of aluminum oxide is readily explained. The trivalent ion has a noble gas configuration, hence a comparatively large amount of energy is required for the removal of the fourth electron. Aluminum occupies a titanium site in the crystal structure but remains in the trivalent state. The association of trivalent aluminum and pentavalent niobium ions preserves electrical neutrality without the formation of trivalent titanium ions. Hence the large reduction in the electrical conductivity of the niobium-doped rutile resulting from the addition of aluminum oxide.

Unfortunately this explanation does not account for the enormous difference between the effects of the other oxides listed in Table IV. The fourth ionization potentials are believed to lie in the range 50 to 64 ev (compare with that of aluminum: 120 ev).

The effect is not closely related to the ionic radius, since the value for trivalent gallium (0.26 A) is only slightly smaller than the ionic radius of trivalent chromium (0.64 A).

Gallium, indium, and chromium are known to exhibit more than one valence in the formation of oxides. The extreme values in the table are for the oxides of aluminum and yttrium, both of which form only trivalent ions with noble gas configurations. A satisfactory theory of these effects would have to explain the remarkable difference between the effects of these oxides.

6.4.7. *The Results of Hauffe and Co-Workers*

Hauffe[47] and co-workers have found that the admixture of a small amount of WO_3 to rutile increases the conductivity enormously (page 136 of reference 47). The effect is more pronounced at 600°C than at 900°C. These workers have also investigated the effect of admixture of the oxides of nickel, gallium, aluminum, and chromium on the electrical conductivity of rutile ceramic in the range 600 to 900°C. The effect of the oxides of aluminum, nickel, and gallium was small, and produced no apparent change in the activation energy for the impurity range 0 to 2.0 mole percent.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

TABLE IV. Effect of 1 mole % of (other) impurity on rutile containing 0.5 mole % Nb_2O_5 ($\sigma = 3\times10^{-4}$ ohm^{-1} cm^{-1}) (after Johnson[a]).

Impurity	Change in conductivity
Al_2O_3	÷35 000
Y_2O_3	÷ 58
Ga_2O_3	÷80 000
In_2O_3	÷ 990
Cr_2O_3	÷ 550

[a] See reference 55.

6.4.8. *Effect of Impurity Concentration*

Johnson attempted to estimate the percentage of foreign oxide that can be taken into solution at 1200°C and retained in this state during subsequent heat treatment, i.e., air-quenching to 250°C. This result showed a saturation effect in most cases at concentration above about 0.25 to 0.50 mole percent.

The conductivity was highly dependent on the heat treatment a sample received, and the rate of cooling determined the degree to which "exsolution" occurred. It would be of interest to establish the extent to which these effects are (a) dependent on the formation of impurity bands and (b) are determined by the solubility of the added oxide in rutile. In case (a) it is probable that there exists a correlation between these results and the paramagnetic susceptibility of these doped specimens.

7. THERMOPOWER

7.1. Introduction

The electrical circuit for the measurement of the thermopower (or Seebeck coefficient) of a semiconductor S with respect to a metal M, is represented schematically in Fig. 6. The junctions are maintained at temperatures T_1 and T_2, where T_2 is greater than T_1. The thermopower of S with respect to M is defined as

$$Q_S - Q_M = -d(V_2 - V_1)/d(T_2 - T_1) \qquad (7\text{-}1)$$

when the net current in the circuit is zero.

Since thermal emf's are additive, it has a meaning to talk of the absolute thermopower of a substance. This is particularly true of semiconductors, since they exhibit much larger thermoelectric effects than metals.

It is shown by Seitz[57] and by Domenicali[58] that the thermopower, Q, consists of the two terms,** and can be represented by

$$Q = (1/e)((\eta - Q^*)/T). \qquad (7\text{-}2)$$

Q^* is the heat transport per electron, which in semiconductors is approximately $2kT$, and so this term

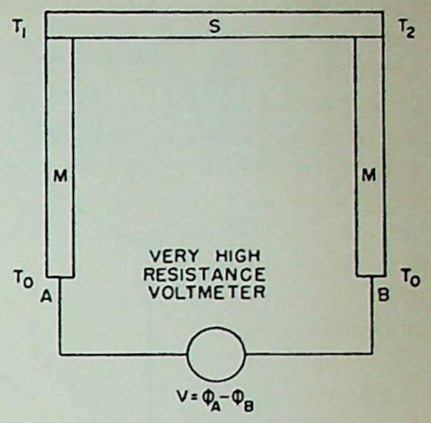


FIG. 6. Circuit for the measurement of the thermopower of a semiconductor S with respect to a metal M, using an electrostatic voltmeter V.

contributes about -170 μv per degree to the thermopower, at all temperatures. The first term on the right-hand side of Eq. (7-2) has the same sign as the Fermi potential of the electrons, η, so that when the Fermi level lies below the bottom of the conduction band this term is negative.

In practice, the small thermopower of the metal is usually neglected in comparison with the much larger value for the semiconductor, so that the measured potential difference is given by the expression

$$F_s = V_1 - V_2 = \int_{T_2}^{T_1} Q_S dT. \qquad (7\text{-}3)$$

The interpretation of Q_S in terms of the choice of band model is discussed in Sec. (5-5).

7.2. Experimental Results

Kataoka and Suzuki[59] have measured the thermopower of ceramic rutile in various stages of reduction. Their high resistivity specimens (3000 ohm cm at 300°K) exhibited a thermopower of -1.03 mv/°C in the range 0 to 90°C. The thermopower, as expected, decreased with increasing impurity content. The specimens of lowest resistivity (0.05 ohm cm at 300°K) exhibited a thermopower of -0.17 mv/°C between 0 and 100°C, falling to about -0.15 mv/°C between 150 and 200°C. Some of their curves of thermopower *versus* temperature were concave toward the temperature axis.

These workers claim to have established a correlation between thermopower and the conductivity of the specimens. The difficulties encountered in making such a correlation due to the uncertainty in the values of the mobility, the effective mass, and in the dynamic term Q^*/eT of the thermopower are discussed in Sec. (5.5.2).

However, the thermopower may be plotted against reciprocal temperature to yield a value for the ioni-

[57] Frederick Seitz, *The Modern Theory of Solids* (McGraw-Hill Book Company, Inc., New York, 1940).

[58] C. A. Domenicali, Revs. Modern Phys. **26**, 237 (1954).

** It is assumed that these transport phenomena are determined entirely by the electronic mobility, and not by that of holes.

[59] S. Katoaka and T. Suzuki, Bull. Electrotech. Lab. Tokyo **18**, No. 1, 732 (1954).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

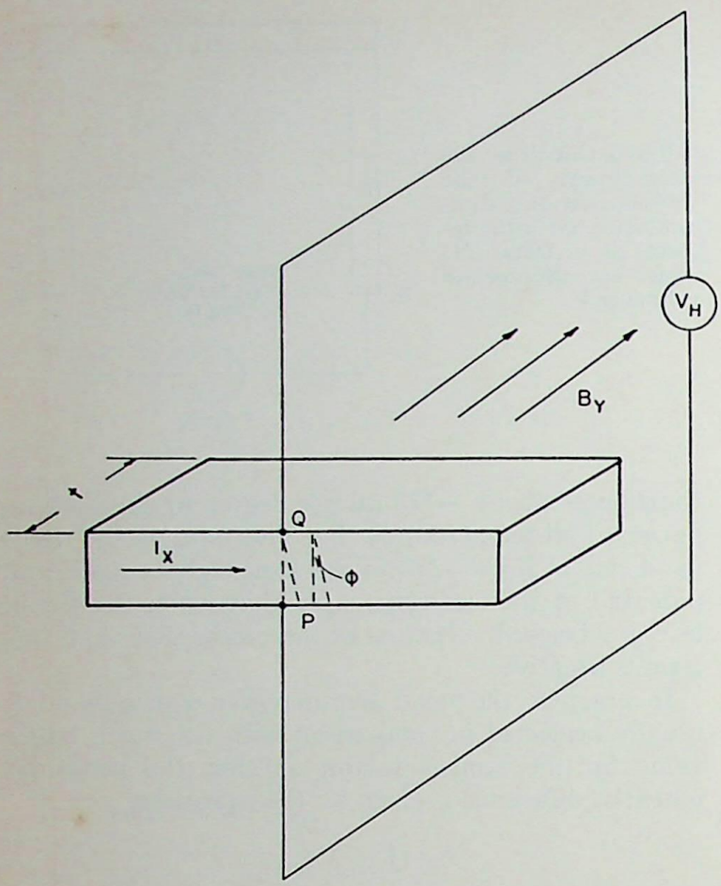


FIG. 7. Schematic diagram of apparatus for Hall effect measurements.

zation potential of the donor sites, to the extent that the simple band model used in deriving Eq. (5-23) is valid. This procedure, applied to the data on page 735 of reference 59 yields values of E_d (noninteracting donor model) in the range -0.02 to -0.13 ev. There appears to be no correlation between the values obtained in this way and those obtained from conductivity data, discussed in Sec. (6.3.1).

When the values of the Fermi level (donor band model) are obtained from these same data by the use of Eq. (5-24), one obtains values ranging from -0.31 ev to zero, the position of the Fermi level rising with increasing donor concentration (as evidenced by the resistivity measurements).

These workers have also shown that the thermopower of their specimens is dependent on the ambient atmosphere. At a pressure of 10^{-1} of mercury the thermopower of their specimens was more highly temperature dependent than at atmospheric pressure.

Boltaks and co-workers[60] have investigated the thermopower Q of sintered rutile that had been reduced in hydrogen or in carbon monoxide. In all cases the thermal emf was negative. Here, Q was found to decrease with increased reduction. The experimental values lay in the range -250 to -415 μv per degree. Thus, Q decreased with increasing temperature for the slightly reduced samples. The thermopower of the highly reduced samples, however, increased with increasing temperature. This latter effect may have resulted from partial reoxidation.

[60] Boltaks, Vasenin, and Salunina, Zhur. Tekh. Fiz. 21, 532 (1951).

8. THE HALL EFFECT

The Hall effect arises from the force experienced by a charge moving in a magnetic field. A high resistance voltmeter is connected to the electrodes P and Q on the lower and upper faces (Fig. 7) of the specimen of uniform rectangular cross section, and thickness t cm, in which a current I_x amp is flowing. The electrodes P and Q are assumed to be placed so that they are at the same potential before the application of a magnetic field. When a magnetic field B_y gauss is applied in a direction normal to the current I_x and parallel to the t direction, a potential difference $V_H(=V_Q-V_P)$ volts appears between the electrodes. The quantity $R=10^8V_Ht/B_yI_x$ is defined as the Hall coefficient. When lattice scattering predominates, the concentration of charge carriers n is related to the Hall coefficient by the formula

$$n=-7.40\times10^{18}G/R. \tag{8-1}$$

For n-type semiconductors, G is a function of the Fermi level which varies between 1 for nondegenerate samples, to $8/3\pi$ for completely degenerate samples.[61,62] When impurity scattering cannot be neglected, this value of n must be corrected[63] by a factor r, which has a value in the range 1 to 2 for ionized impurity scattering.

8.1. Hall Data of Reduced Rutile

Breckenridge and Hosler[61] have made Hall measurements on samples of rutile ceramic and single crystals which had been subjected to various degrees of reduction in hydrogen. These workers then attempted to interpret the values of the electron concentration n obtained by the use of Eq. (8-1) in terms of two sources of conduction electrons, one of which has a very small activation energy, so that it was considerably ionized even at liquid air temperature. Their results (Fig. 8) were interpreted in terms of the following equation

$$n=n_1\exp(\Delta E_1/2kT)+n_2\exp(\Delta E_2/2kT) \tag{8-2}$$

with $\Delta E_1\ll\Delta E_2$, and $n_1\ll n_2$.

Interpretation of ΔE_1 and ΔE_2 in terms of the ionization potentials of isolated donor type impurities is complicated by the effect of the $\frac{3}{4}\ln t$ correction term of Eq. (5-6). This correction is important even for ΔE_2, whereas the significance of the smaller ΔE_1 is almost entirely lost because of the magnitude of the correction term.

[61] R. G. Breckenridge and W. R. Hosler, Phys. Rev. 91, 793 (1953).

[62] K. Shifrin, J. Phys. (U.S.S.R.) 8, 242 (1944).

[63] W. Crawford Dunlap, Jr., *An Introduction to Semiconductors* (John Wiley & Sons, Inc., New York, 1957); see Fig. 6.5.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

Attempts have been made below to interpret the Hall data of Breckenridge and Hosler[61] under the assumption that the form of the curves at higher temperatures (200 to 500°K) is determined by the electrons in the conduction band, to which Maxwell-Boltzmann statistics and the concept of a single effective electron mass are assumed to apply. In the first case, the source of these electrons is considered to be non-interacting donor sites at an energy E_d with respect to the bottom of the conduction band. The appropriate type of plot for the experimental data is suggested by Eq. (5-6), and from the slope a value of E_d may be obtained.

In the second case, the appropriate type of plot is indicated by Eq. (5-7), in which no assumption concerning the source of electrons is implied. However, the slope of the plot yields the height of the Fermi level, which will lie near the center of the donor band if it exists. In this case the plot yields also the effective mass of the conduction electrons.

At the temperatures considered, the effect of the magnetic field on the concentration of conduction-band electrons is believed to be entirely negligible.[64]

When the data of Breckenridge and Hosler's Fig. 11 are plotted in these two different ways for the temperature range 250 to 500°K, the results are as follows. With the exception of one very low value, for curve (*b*), the values of E_d (noninteracting donors) lie in the range -0.095 to -0.147 ev. With the same exception, the values of the Fermi level (donor band model) lie in the range -0.025 to -0.051 ev with respect to the bottom of the conduction band. The corresponding values of the effective mass ratio are in the range 0.16 to 1.6. Although Maxwell-Boltzmann statistics can scarcely be said to apply to these data, these values of the effective mass of the conduction electrons are much lower than the values usually quoted for rutile.

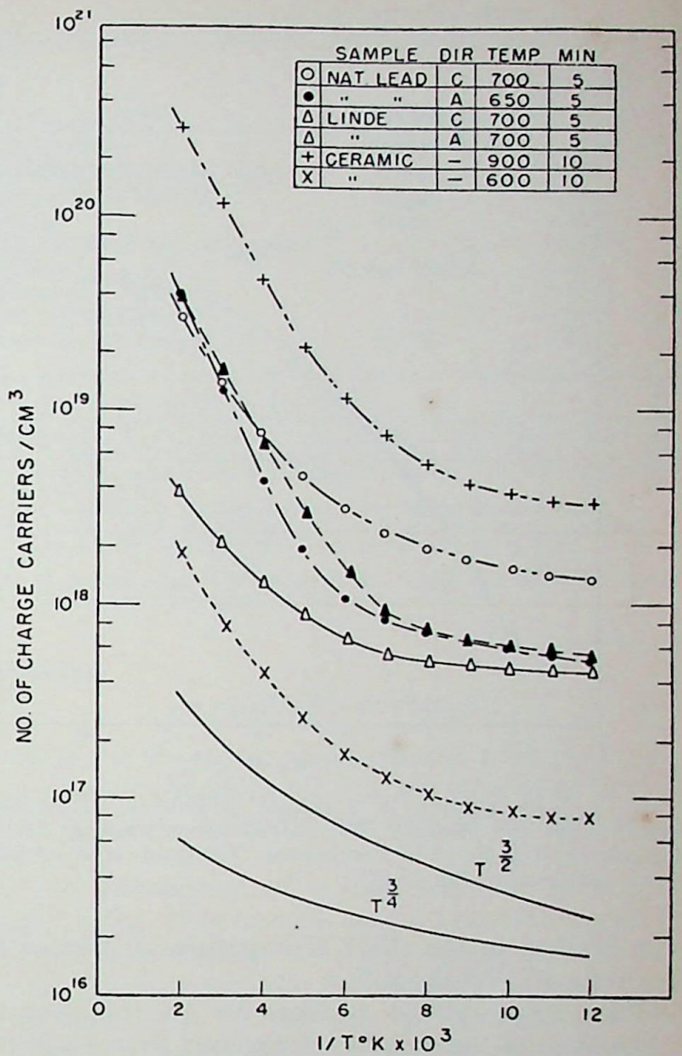


FIG. 8. Electron concentrations in specimens of rutile reduced in hydrogen under the conditions specified (after Breckenridge and Hosler[61]). Curves showing the temperature variation of the factors $T^{3/4}$ and $T^{3/2}$ of Eqs. (5-5) and (5-7) are shown for comparison.

8.2. Hall Mobility

An electric charge moving in the x direction normal to a magnetic field B_y, "sees," in addition, an electric field E_z, where

$$E_z = v_x B_y \quad \text{(mks units)}.$$

Since the mobility b is defined as v_x/E_x, we obtain

$$E_z/E_x = b_H B_y \quad \text{(mks units)}.$$

Hence the equipotential surfaces in a conductor are rotated by the magnetic field through an angle ϕ, where $\tan\phi = b_H B_y$. In cgs units

$$\tan\phi = 10^{-8} b_H B_y.$$

This simple relation is somewhat obscured by the usual derivation of the Hall mobility from the electrical conductivity and the Hall coefficient

$$\sigma = neb$$
$$-R = 1/ne$$
$$(-R\sigma) = b_H. \qquad (8\text{-}3)$$

The determination of b_H from simultaneous measurements of σ and R is entirely equivalent experimentally to the measurement of $\tan\phi$, so that the simple interpretation of the Hall mobility is unaltered.

Figure 9 illustrates a typical Hall mobility curve derived from the electrical conductivity and Hall data of Breckenridge and Hosler. These workers attempted to interpret their results in terms of the theory of scattering by optical lattice vibrations developed by Fröhlich and Mott,[65] and by Fröhlich, Pelzer, and Zineau.[66] The additional scattering at low temperatures

[64] P. T. Landsberg, Proc. Phys. Soc. (London) 71, 69 (1958).

[65] H. Fröhlich and N. F. Mott, Proc. Roy. Soc. (London) A171, 496 (1939).

[66] Fröhlich, Pelzer, and Zineau, Phil. Mag. 41, 221 (1950).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

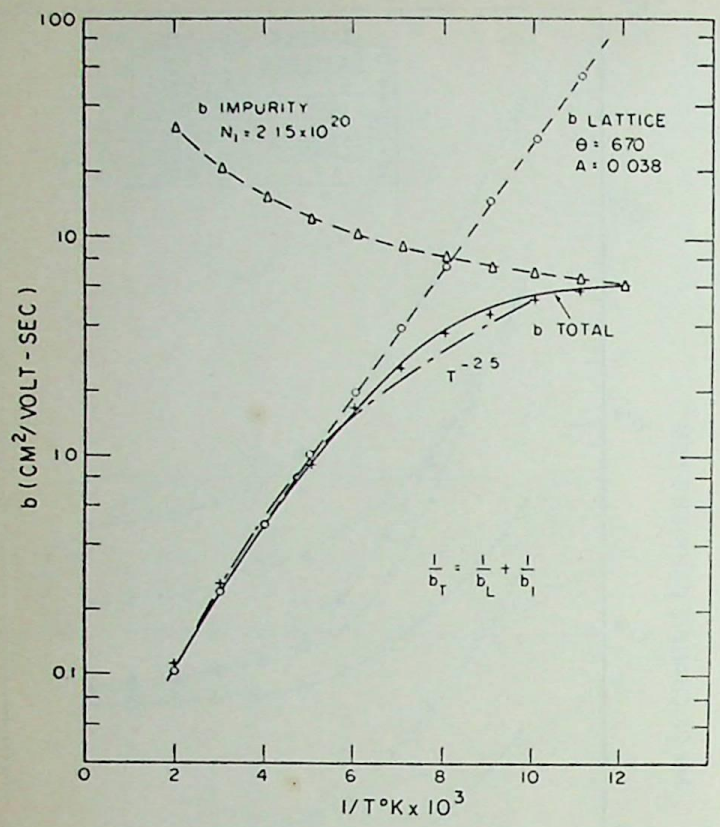


FIG. 9. Hall mobility in a sample of reduced rutile, after Breckenridge and Hosler.[61] The dotted curve, showing a $T^{-2.5}$ dependence, is included for comparison. The symbols and A refer to the constants in Eq. (5-13).

was ascribed to the effect of impurities as treated by Conwell and Weisskopf.[67]

When an empirical relation for the temperature dependence of the mobility is required to assist in the interpretation of conductivity data above about 200°K, a $T^{-2.5}$ law appears to be satisfactory. For comparison purposes, a $T^{-2.5}$ plot is included in Fig. 9.

8.3. Magnetoresistance

Gorelik[68] has investigated the effect of a magnetic field of 15 000 gauss on the electrical conductivity of "pure" rutile, and on specimens of rutile containing various proportions of the oxides of calcium and strontium. The entire change in conductivity did not occur immediately after the application of the magnetic field, but was practically complete after two minutes. The change in conductivity was small at room temperatures, but above about 160°C the fractional change in conductivity rose to steady values of one or two percent.

Gorelik has given an "explanation" of these effects in terms of the relative importance of the ionic and electronic contributions to the conductivity.

9. MAGNETIC SUSCEPTIBILITY

9.1. Introduction

In a semiconductor, such as rutile, one must consider the following four contributions to the magnetic susceptibility[69]: (a) the lattice susceptibility, (b) the carrier susceptibility, (c) the contribution of the impurities, including effect of departures from stoichiometry. This effect may be differentiated into two cases (i) the effect of noninteracting sites, and (ii) the effect of donor (or acceptor) bands; and (d) the effect of surface states, including adsorbed gases.

9.2.(a) Lattice Susceptibility

Van Vleck[70] gives the following expression for the lattice susceptibility, and points out that there is no sharp dividing line between diamagnetic molecules and feebly paramagnetic ones:

$$\chi_{\text{mole}}=-\frac{Le^2}{6mc^2}\sum\langle r^2\rangle+\tfrac{2}{3}L\sum_{n'\neq n}\frac{|m^0(n';n)|^2}{h\nu(n';n)}. \quad (9\text{-}1)$$

In Eq. (9-1), $m^0(n';n)$ is a nondiagonal element of the vector matrix for the magnetic moment of the system, evaluated in the absence of a field, $\nu(n';n)$ is the frequency corresponding to the $n';n$ transition. L is Avogadro's number, e the electronic charge in esu, m the electronic mass in grams, and c is the speed of light in cm/sec. $\sum\langle r^2\rangle$ is the mean value of the radius of an electron orbit, summed over all the electrons.

When the second term of Eq. (9-1) predominates, the contribution represents a small, almost temperature-independent paramagnetism. Using the simple ionic model, the ions in the oxides Sc_2O_3, TiO_2, V_2O_5, and CeO_2 possess noble gas configurations, hence would be expected to have a negative (diamagnetic) lattice susceptibility. However, there appears to be some evidence, especially in the case of V_2O_5, for a contribution due to temperature-independent paramagnetism larger than the negative (diamagnetic) term.[71–74] Freed and Kasper[75] discuss this effect in ions of the type VO_2^{+2}, WO_4^{-2}, MnO_4^-, VO_3^-, and $Cr_2O_7^{-2}$. Bates[76] considers that the feeble paramagnetism believed to exist in Sc_2O_3, TiO_2, and CeO_2 may be caused by the "distortion due to interatomic forces," and adds that it is difficult to determine in a given case whether the effect is due to a negative exchange integral, to electron migration (in conductors), or to ionic deformation.

At least in the case of rutile, the wide range of experimental values reported in the literature, which include both positive and negative values, may result from departures from stoichiometry. The formation of oxygen vacancies (and Ti^{+3} sites) which result when

[67] E. Conwell and V. F. Weisskopf, Phys. Rev. 77, 388 (1950).
[68] S. I. Gorelik, J. Exptl. Theoret. Phys. 21, 826 (1951).
[69] Stevens, Cleland, Crawford, and Schweinler, Phys. Rev. 100, 1084 (1955).
[70] J. H. Van Vleck, *The Theory of Electric and Magnetic Susceptibilities* (Oxford University Press, New York, 1932), pp. 275, 279, 302–303.
[71] T. Ishiwara, Sci. Rept. Tôhoku Univ. 3, 303 (1914).
[72] P. Weiss and P. Collet, Compt. rend. 178, 2146 (1924); 181, 1057 (1925).
[73] R. Ladenburg, Z. physik. Chem. 126, 133 (1927).
[74] L. F. Bates, *Modern Magnetism* (Cambridge University Press, New York, 1951), pp. 272, 44.
[75] S. Freed and C. Kasper, J. Am. Chem. Soc. 52, 4671 (1930).
[76] Reference 74, pp. 281 and 324.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

rutile is prepared at moderately high temperatures in a reducing atmosphere is discussed in Sec. 14. The various values of the magnetic susceptibility of "pure, stoichiometric" rutile, appearing in the literature, are listed in Table V.

9.3.(b) Carrier Susceptibility

Assuming that the carrier electrons behave as a perfect electron gas, and using the appropriate density-of-states effective mass ratio, the following expression for the carrier susceptibility has been derived[69,77]:

$$\chi_c=\left(\frac{\beta^2}{3kT}\right)n_c[3-\langle f_e^2\rangle_{Av}], \tag{9-2}$$

where n_c is the concentration of electrons in the conduction band, β is the Bohr magneton, and $\langle f_e^2\rangle_{Av}$ depends on the curvature of the energy surfaces in k space. This term is related to the effective mass ratio for orbital motion M^M by the relation

$$M^M=(\langle f_e^2\rangle_{Av})^{-\frac{1}{2}}. \tag{9-3}$$

Since M^M is expected to be considerably larger than unity in the case of rutile,[78] Eq. (9-2) may be approximated by the relation

$$\chi_c=\beta^2 n_c/kT, \tag{9-4}$$

which is of the same form as the contribution of unpaired electrons in noninteracting hydrogen-like donors to be discussed in the next paragraph.

9.4.(c) Impurity Contribution

Mooser[79] has discussed (i) the contribution of electrons in noninteracting hydrogen-like impurity sites, and (ii) the case in which the overlapping of the impurity wave functions is sufficiently great for the formation of impurity bands. Considering first the case in which the donor sites are noninteracting and are in s-like ground states, he obtains

$$\chi_{\text{para}}=n_d\beta^2/kT, \tag{9-5}$$

where n_d is the concentration of occupied donors, which is related to N_d, the concentration of donor sites, by the following relation:

$$n_d=N_d/1+\tfrac{1}{2}\exp((E_d-\eta)/kT). \tag{9-6}$$

Here, $E_d-\eta$ is the energy of the donor levels with respect to the Fermi level η.

A diamagnetic term must be added to Eq. (9-5) so that the total contribution of the electrons in isolated, hydrogen-like donor sites is given by the following expression:

$$\chi=n_d(\beta^2/kT-e^2r^2/6m_ec^2) \tag{9-7}$$

in which r is the radius of the electron orbit in cm, e is the electronic charge in esu, c is the velocity of light in cm sec^{-1}, and m_e is the effective mass of the electrons in grams.

If the assumption of a high effective mass M^M, implied in Eq. (9-4) is valid, comparison of this relation with the expression for the paramagnetic contribution of (atomic) hydrogen-like donors shows that they are of the same form. In this case, the paramagnetic susceptibility is independent of the distribution of the unpaired electrons between the donor sites and the conduction band, i.e., the thermal ionization of donor sites does not affect their contribution to the susceptibility. This is in contrast to the case in which the wave functions of the impurity sites overlap sufficiently for the formation of impurity bands. This point is discussed later.

TABLE V. Magnetic susceptibility of rutile.

Source	Temperature	Magnetic susceptibility (emu per gram)
Meyer[a]	288°K	0.37×10^{-6}
Wedekind and Hausknecht[b]	room	0.066
Berkman and Zocher[c]	room	−0.20
Hüttig[d]	room	−0.30
Raychaudhuri and Sengupta[e]		0.073
Ehrlich[f]	90 and 293°K	0.08
Zimens and Hedvall[g]		0.134
Hill and Selwood[h]	[i]	−0.3
Reyerson and Honig[j]	301°K	0.061[k]

[a] S. Meyer, Ann. Physik 69, 236 (1899).
[b] E. Wedekind and P. Hausknecht, Ber. deut. chem. Ges. 46, 3763 (1913).
[c] S. Berkman and H. Zocher, Z. physik. Chem. 124, 322 (1926).
[d] G. F. Hüttig, Z. anorg. u. allgem. Chem. 224, 225 (1935).
[e] D. P. Raychaudhuri and P. N. Sengupta, Indian J. Phys. 10, 253 (1936).
[f] See reference 80.
[g] K. Zimens and J. Hedvall, Svensk Kem. Tidsk. 52, 12 (1941).
[h] F. N. Hill and P. W. Selwood, J. Am. Chem. Soc. 71, 2522 (1949).
[i] "Substantially independent of temperature," presumably down to −190°C.
[j] L. H. Reyerson and J. M. Honig, J. Am. Chem. Soc. 75, 3920 (1953).
[k] See discussion of adsorption of NO_2-N_2O_4 on rutile.

9.4.1. *Electrons in Impurity Bands*

The diamagnetic contribution of electrons in impurity bands is believed to be small and about the same as that of electrons in noninteracting donor sites [the second term in Eq. (9-7)].

Band wide compared with kT.—At low temperatures the paramagnetic contribution is given by the following expression

$$\chi_{\text{para}}=2\beta^2D(\eta) \tag{9-8}$$

in which $D(\eta)$ denotes the density of states in the donor band at the position of the Fermi level. In this temperature range χ should be almost independent of temperature.

Band narrow compared with kT.—At high temperatures, where the band width is much less than kT, the paramagnetic contribution of the electrons in the

[77] H. Fröhlich, *Theorie der Metalle* (Verlag Julius Springer, Berlin, 1937), pp. 144–156.
[78] R. G. Breckenridge and W. R. Hosler, Phys. Rev. 91, 793 (1953).
[79] E. Mooser, Phys. Rev. 100, 1589 (1955); G. Busch and E. Mooser, Helv. Phys. Acta 24, 329 (1951); 26, 611 (1953); Z. physik. Chem. 198, 23 (1951).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

TABLE VI. Fraction of donor electrons contributing to magnetic susceptibility of TiO_x (after Ehrlich).

x	Temperature 3°K	O_v Oxygen vacancies per mole of TiO_2	n_D Donor electrons per mole of TiO_2 (2 per O_v)	Experimental χ_{para} per mole[a]	n_D' (calculated from χ_{para}) donor electrons per mole	$f=\frac{n_D'}{n_D}$
1.97	293	1.8×10^{22}	3.6×10^{22}	46×10^{-6}	2.2×10^{22}	0.61
1.95	293	3.0	6.0	72	3.5	0.57
1.90	293	6.0	12.0	119	5.7	0.47
1.97	90	1.8	3.6	104	1.54	0.42
1.95	90	3.0	6.0	162	2.39	0.40
1.90	90	6.0	12.0	216	3.19	0.26

[a] After correction for temperature-independent paramagnetic lattice susceptibility ($+6\times10^{-6}$ emu per mole).

impurity band is obtained from the result obtained for the noninteracting donor sites by multiplying by the factor f, which indicates the fraction of the electron spins that are "free" (see Sec. 9.4.2)

$$f=\exp((\bar{E}_d-\eta)/kT)/(1+\exp(\bar{E}_d-\eta)kT), \quad (9\text{-}9)$$

where $\bar{E}_d$ is the energy of the midpoint of the donor band. This factor has the value unity when the Fermi level lies far below the donor band. In the case of principal interest $(\bar{E}_d-\eta)$ is zero or slightly negative, and the multiplying factor is approximately equal to $\frac{1}{2}$. When the Fermi level is more than $3kT$ higher than the donor band f is very small, tending to zero with increased separation.

The numerical value of this factor is therefore highly dependent on the height of the Fermi level. The dependence of the Fermi level on temperature has been discussed in Sec. 5.

Once again it is necessary to draw attention to the high dielectric constant in rutile. It is therefore probable that impurity bands form at lower impurity concentrations than in most other materials. This effect may be offset by the possibly high effective electron mass.

9.4.2. Donor Sites in Nonstoichiometric Rutile

Although the contribution of the Ti^{+4} ions to the magnetic susceptibility is small, and of uncertain sign, the Ti^{+3} ion definitely makes a large contribution to the paramagnetic susceptibility of the crystal. Departures from stoichiometry of rutile are believed to result in the conversion of Ti^{+4} sites to Ti^{+3} sites. These ions of lower valence may in many ways be considered to constitute an "impurity" in the rutile lattice, and in this section the terms "impurity site" and "impurity band" are understood to apply to nonstoichiometric rutile.

Ehrlich[80] has investigated the magnetic susceptibility of solids of the composition TiO_x from $x=0$ to $x=2$. If it is assumed that the formation of an oxygen vacancy results in the formation of two Ti^{+3} sites, each with a magnetic moment of 1.73 Bohr magnetons, one may calculate the resulting paramagnetic susceptibility. For example, Ehrlich estimated that the paramagnetic susceptibility of rutile which had been reduced to the extent of two percent weight loss ($x=1.90$) should be 253×10^{-6} per gram mole at room temperature. Since the experimental value was 119×10^{-6} he concluded that only 47% of the electrons were "free."

Ehrlich's calculations have been extended to his other data, and are summarized in Table VI. As expected, the fraction f of the donor electrons that contribute to the susceptibility (column 7) decreases with increasing donor concentration.

However, we note the higher values of f for the room temperature measurements as compared with the measurements at 90°K.

Now subject to the assumptions of a high effective mass ratio M^M and the applicability of Maxwell-Boltzmann statistics, the raising of electrons from isolated hydrogen atom-like donors to the conduction band should not appreciably alter the paramagnetic susceptibility. We therefore look for an explanation in terms of overlapping of donor wave functions and the formation of a donor band.

A very approximate value for the width of this hypothetical band may be obtained from Ehrlich's low temperature data by the application of Mooser's expression for the susceptibility of electrons in bands wide in comparison with kT. Here, $\chi=2\beta^2D(\eta)$ where $D(\eta)$ represents the density of states at the Fermi level. The total number of levels was taken to be twice the number of oxygen vacancies. The approximation was introduced that the density of states D was uniform throughout the band, whence ΔE was estimated. The results of calculations for TiO_x (where $x=1.90$, 1.95, and 1.97) at 90°K appear in Table VII. The estimated values of ΔE are somewhat larger than kT (which is 0.0077 eV at 90°K), so that the requirement regarding the width of the band is fulfilled.

It is now necessary to account for the observed increase in f with increasing temperature. One may ask whether this increase could reflect a transition from the "wide band" case, at 90°K, to the "narrow band" case. The estimated band widths (last column of Table VII) are about fifty percent greater than kT at 293°K. It is not unreasonable to suppose that the 293° values represent an intermediate case, and that at still higher

[80] P. Ehrlich, Z. Electrochem. 45, 362 (1939).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

temperatures the "narrow band" approximation would be valid. Earlier, it was concluded that the Fermi level would be expected to lie at, or somewhat lower than, the middle of the donor band. In the "narrow band" limit, the expected value of f calculated from Eq. (9-8) is therefore equal to, or greater than, $\frac{1}{2}$. The observed values of f (given in the last column of Table VI) appear to be consistent with this interpretation.

An alternative explanation may be given if two electrons are bound to each oxygen vacancy at sufficiently low temperatures, and these sites are assumed to be noninteracting. One would expect the electron spins to be paired in such a "molecule." Raising of one or both of these electrons to higher states by thermal excitation would uncouple the spins. This conclusion is valid, regardless of the extent of the overlapping of the wave functions of the higher states, provided that the Fermi level lies farther than about $3kT$ below the bottom of any such higher band [see Eq. (9-9)]. If the Fermi level lay closer to this higher band than about $3kT$, the Pauli principle, would have to be invoked, and the value of f would remain below unity.

Ehrlich's results may also be expressed in terms of the Curie-Weiss law (discussed further in the next section), in which case one obtains an effective magnetic moment of 1.5 Bohr magnetons, instead of the expected value of 1.73, and Weiss constants of 700, 72, and 160°K, respectively, for the three compositions $x=1.97$, 1.95, and 1.90.

Gray[81] has reported an investigation of the magnetic susceptibility of "titanium dioxide" during reduction by hydrogen. Simultaneous measurements were made of the decrease in hydrogen pressure, sample weight loss, and paramagnetic susceptibility increase, the temperature being maintained at 560°C. There appears to be an almost linear relation between change in susceptibility and the loss in weight. The sample weight and the method of measurement were not stated and without this information a satisfactory interpretation of the data cannot be made.

9.4.3. *Rutile Containing Impurities*

Representation of the paramagnetic susceptibility arising from donor sites in solids is not often made in terms of the foregoing model invoking donor bands. More commonly the departures from the Curie law are represented empirically by replacement of the T factor in this law by the factor $(T-\Delta)$. This procedure is quite empirical, and may include the effects resulting from the "pairing" of electrons, and the formation of donor bands, discussed in the foregoing.[82,83] Most of the published experimental results are expressed in terms of the magnetic moment of the unpaired electron and the corresponding Weiss constant Δ.

TABLE VII. Width of hypothetical impurity band in TiO_x estimated from Ehrlich's 90°K magnetic susceptibility data.

x	χ_{para} per mole[a]	Number of states per mole (2 per O_a)	Calculated density of states at Fermi level $D(\eta)$ per ev	ΔE estimated width of donor band ev
1.97	104×10^{-6}	3.6×10^{22}	99×10^{22}	0.036
1.95	162	6.0	154	0.039
1.90	216	12.0	206	0.058

[a] After correction for temperature-independent paramagnetic lattice susceptibility ($+6\times10^{-6}$ emu per mole).

Selwood[84] and co-workers have shown the "pure" titanium sesquioxide to be antiferromagnetic, with a critical point near the specific heat anomaly at about 248°K. The Weiss constant was unusually large, and was estimated to be about 2000°K. The corresponding magnetic moment for the Ti^{+3} ion was 1.2 Bohr magnetons, which may be compared with the expected value of 1.73 for the free Ti^{+3} ion with one unpaired electron. These results are contradicted by the later work of Foëx and Wucher,[85] and of Pearson.[86] These workers found a continuous increase in susceptibility as the temperature was increased from −200 to about 360°C, the temperature dependence being most pronounced in the vicinity of 160°C.

When paramagnetic ions are separated by solution in a suitable diamagnetic crystal, their mutual interaction is reduced. An example of this effect appears in the work of Adler and Selwood,[87,88] who have studied a solid solution of 14.3% Ti_2O_3 in the diamagnetic oxide, Al_2O_3. The Ti^{3+} ion was found to have a magnetic moment of 1.1 Bohr magnetons (expected value $\sqrt{3}$ Bohr magnetons), and a value of $\Delta=105$°K was obtained. This value of the Weiss constant Δ may be compared with the value of about 2000° obtained for undiluted Ti_2O_3.

Selwood and co-workers have investigated the incorporation of ions of other transition metals into the rutile crystal lattice. The Weiss constant of the manganese ion tends to zero as expected at infinite dilution, and is almost proportional to the concentration.[89] At a concentration of 19% manganese by weight, Δ is about 315°. The magnetic moment corresponds to a valence state of about 4.1 at zero concentration, falling to

[81] T. J. Gray, *The Defect Solid State* (Interscience Publishers, Inc., New York), p. 287.

[82] Other factors affecting the Weiss constant: the multiplet separations and the effect of inhomogeneous electric crystalline fields are discussed on page 150 of reference 84.

[83] R. K. Wangness has discussed the relation between the exchange coupling between spins, and the Curie-Weiss law. R. K. Wangness, J. Chem. Phys. **20**, 1656 (1952).

[84] Pierce W. Selwood, *Magnetochemistry* (Interscience Publishers, Inc., New York, 1956), p. 329.

[85] M. Foëx and J. Wucher, Compt. rend. **241**, 184 (1955).

[86] A. D. Pearson, Technical Report 120, Laboratory for Insulation Research, Massachusetts Institute of Technology, (1957).

[87] Reference 84, p. 161.

[88] S. F. Adler and P. W. Selwood, J. Am. Chem. Soc. **76**, 346 (1954).

[89] Selwood, Moore, Ellis, and Wethington, J. Am. Chem. Soc. **71**, 693 (1949).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

about 3.9 as the concentration is increased to 19% manganese (by weight).

The valence state of nickel incorporated into the rutile lattice decreases somewhat with increasing concentration.[90] Extrapolation back from the lowest concentration appearing in the curves (about four percent nickel by weight) indicates that the magnetic moment at infinite dilution is probably between 3.8 and 4.0, considerably lower than expected for six (or four unpaired) electrons. However, from the color, and for other reasons, Selwood believes that the oxidation state is indeed +4.

Similar results were obtained for chromium[91] and iron[92] incorporated into the rutile lattice. In the case of iron, Δ fell from 52° at 12.9% iron, to 16° at 1.5% iron (by weight). The corresponding values of the oxidation state were 3.2 and 3.9. The chromium ion in rutile (2.6% by weight) exhibited a magnetic moment of 2.8 Bohr magnetons, and a negative Δ (−27°).

Thus with the possible exception of nickel, these elements assume a valence state of almost four when occupying titanium sites (at sufficient dilution). Selwood has applied the term "induced valence" to this phenomenon.

9.5.(d) Surface Contribution

The magnetic susceptibility will include the contribution of adsorbed gases, in particular physically adsorbed oxygen. The meaning of the term chemisorption when applied to oxygen and metallic oxides is not clear. However, the nature of the space-charge-barrier layer at the surface of a semiconductor (of the order of 10^{-5} cm thick) is influenced by the chemisorption of gases. Any factor which influences the height of the Fermi level in this region will alter the paramagnetic susceptibility. The importance of this effect is expected to increase as the particle size is reduced.

Sandler[93] has used the rate of *ortho-para* conversion of hydrogen on rutile powder as an indication of the surface density of unpaired electrons arising from the paramagnetism due to surface defects. These defects are believed to be due to a stoichiometric deficiency of oxygen. This surface paramagnetism is apparently associated with the blue color of reduced rutile.

A mention of the investigation of the surface paramagnetism of rutile conducted by Reyerson and Honig, and of the effect of reoxidation by NO_2-N_2O_4 will be found in Sec. 12.

[90] F. N. Hill and P. W. Selwood, J. Am. Chem. Soc. **71**, 2522 (1949).
[91] P. W. Selwood and L. Lyon, Discussions Faraday Soc. **8**, 222 (1950).
[92] Selwood, Ellis, and Wethington, J. Am. Chem. Soc. **71**, 2181 (1949).
[93] Y. L. Sandler, J. Phys. Chem. **58**, 54 (1954).

10. THERMAL CONDUCTIVITY, SPECIFIC HEAT, AND VARIOUS PHYSICAL PROPERTIES

10.1. Thermal Conductivity

Kingery and McQuarrie[94] have determined the thermal conductivity of a number of pure ceramic oxides, including rutile, in the temperature range 100 to 1800°C. The results were corrected for the porosity of the samples. Values of 0.0119 and 0.0081 (cal sec^{-1} cm^{-2}) (°C cm^{-1}) at 200 and 800°C, respectively, were obtained. Kingery[95] has discussed the variations in thermal conductivity associated with the effect of radiant energy transmission, the ratio of the phonon mean free path to lattice dimension, the value of the Debye temperature, emissivity, and the relation of the electronic conductivity to the thermal conductivity of rutile.

The thermal conductivities of sapphire and rutile single crystals have been measured in the temperature range −50 to 100°C by McCarthy *et al.*[96] The thermal conductivity is greater parallel to the optic axis than in the direction perpendicular to it. Both values decrease with increasing temperature in the temperature range studied, and their ratio is also temperature dependent.

10.2. Thermal Coefficient of Expansion

Day[97] has obtained a value of 80×10^{-7} per °C, for the coefficient of expansion of rutile in the range −130 to 50°C. Mauer and Bolz[98] have determined the thermal expansion coefficient of stoichiometric and reduced rutile from x-ray data. In the temperature range 0 to 200°C the average values of the coefficient of expansion of stoichiometric rutile were 8.0×10^{-6} in the a direction and 9.2×10^{-6} per °C in the c direction. In the temperature range 0 to 1000°C the corresponding average values are 8.4×10^{-6} and 9.2×10^{-6} per °C. In the temperature range 0 to 400°C, the average values for reduced rutile ($TiO_{1.97}$) were 8.4×10^{-6} per °C for the a direction, and 10.5×10^{-6} per °C for the c direction.

10.3. Specific Heat and Debye Temperature

At sufficiently low temperatures, only the low-frequency vibrations are important in determining the internal energy of a solid. In this region the specific heat capacity is expected to be proportional to T^3, and one expects a constant value for the Debye temperature θ. In rutile, this range extends to above 4°K, and Keesom and Pearlman[99] obtained a value of 758°K for the corresponding Debye temperature. In the

[94] W. D. Kingery and M. McQuarrie, J. Am. Ceram. Soc. **37**, 107 (1954).
[95] W. D. Kingery, J. Am. Ceram. Soc. **38**, 251 (1955).
[96] McCarthy, Ballard, and Doerner, Phys. Rev. **88**, 153A (1952).
[97] Jean Day, Bull. soc. sci. Bretagne **24**, 13 (1949).
[98] Floyd A. Mauer and Leonard H. Bolz, National Bureau of Standards, Wright Air Development Center Technical Rept. 55-473 (1955).
[99] P. H. Keesom and N. Pearlman, Phys. Rev. **98**, 1539A (1955).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

temperature range 10 to 20°K they found that θ decreases from 650 to 460°K.

Also from specific heat studies, McDonald and Seltz[100] obtained a value of $\theta=670$°K in the temperature range 68 to 298°K.

Basing their calculations on earlier measurements of infrared absorption bands, Breckenridge and Hosler[101] obtained values of $\theta=936$°K and $\theta=1161$°K. However, their interpretation of their Hall mobility data was more consistent with McDonald and Seltz's value of $\theta=670$°K.

After slight reduction in H_2 or Ar, at 1000°C, the heat capacity in the helium region increases greatly.[99] As a tentative explanation, the authors suggested that the additional heat capacity was possibly due to the contribution of the "free" classical electrons. In which band these "free" electrons are to be found, is an open question. From the estimated degeneracy temperature these workers obtained a carrier mass of 350 times the free electron mass.

The heat capacity of rutile powder has been measured in the low temperature region by Shomate.[102] Arthur[103] has made measurements in the range 293 to 1073°K and Naylor[104] in the range 500 to 1800°K. For the heat capacity in calories per mole, valid in the range 500 to 1800°K, Naylor gives the formula:

$$c_p=0.2145+12.3\times10^{-6}T-4380/T^2.$$

Arthur expresses his results in the form

$$c_p=0.1512+0.000245T-0.222\times10^{-7}T^2$$

calories per gram per degree, plus or minus two percent, valid in the range 293 to 1073°K. For purpose of comparison, these formulas yield, respectively, values of 0.217 and 0.333 cal/g/deg, at 800°K.

Dugdale *et al.*[105] expected to find a particle-size effect at low temperatures. Instead, however, they found that the specific heat of rutile was independent of particle size below 50°K. Above this temperature the observed dependence was attributed qualitatively as an effect of the *optical* modes of vibration. Because of the relatively complicated crystal structure of rutile, no attempt was made to make a theoretical calculation of the magnitude of the effect. These workers obtained a value of $C_p=0.115$ cal/mole°K at 20°K and 1.407 cal/mole°K at 50°K.

[100] H. J. McDonald and H. Seltz, J. Am. Chem. Soc. **61**, 2405 (1939).
[101] R. G. Breckenridge and W. R. Hosler, Phys. Rev. **91**, 793 (1953).
[102] C. H. Shomate, J. Am. Chem. Soc. **69**, 218 (1947).
[103] J. S. Arthur, J. Appl. Phys. **21**, 732 (1950).
[104] B. F. Naylor, J. Am. Chem. Soc. **68**, 1077 (1946).
[105] Dugdale, Morrison, and Patterson, Proc. Roy. Soc. (London) A224, 228 (1954).

10.4. Miscellaneous Physical Properties

10.4.1. Density

The density of rutile,[106] calculated from x-ray data is 4.28; the pycnometer density is 4.22 g/cc.[107]

10.4.2. Melting Point

St. Pierre[108] has determined the melting point to be 1840±10°C.

10.4.3. Compressibility

The compressibility of natural rutile single-crystals has been investigated by Bridgman.[109] His results could be expressed up to $p=12\,000$ kg/cm² by the formula

$$\Delta L/L_0=-a_1p+b_1p^2.$$

From 30° to 75°C the value of a_1 for the c direction was found to vary from 1.038 to 1.090×10^{-7}, and in the a direction was found to vary from 1.871 to 1.954×10^{-7}. For all cases investigated, the value of b_1 was 7×10^{-13}.

10.4.4. Hardness

Moore[110] gives a value of 7 to 7½ (Mohs scale) for the hardness of synthetic rutile crystals. This value is slightly higher than that of quartz. Measurements were made on specimens exhibiting various degrees of reduction, but the variation in hardness in these different specimens was about the same as the variation with crystallographic direction.

10.5. Miscellaneous

A list of references for these and other properties of rutile may be found in "Properties of Titanium Compounds and Related Substances," ONR Report ACR-17, October 1956.

11. DIELECTRIC AND OPTICAL PROPERTIES

11.1. Dielectric Constant and Loss Angle

(a) *Stoichiometric rutile.*—The frequency and temperature dependence of the dielectric constant of stoichiometric rutile have been investigated by Eucken and Büchner,[111] Schusterius,[112] Büttner and Engl,[113] Berberich and Bell[114] and by von Hippel and his coworkers.[115] The dielectric constant shows little variation

[106] Skinner *et al.*, "Titanium and its compounds," Herrick L. Johnston Enterprises (1954).
[107] A. Schroeder, Z. Krist. **67**, 485 (1928).
[108] P. D. S. St. Pierre, J. Am. Ceram. Soc. **35**, 188 (1952).
[109] P. W. Bridgman, Am. J. Sci. **15**, 287 (1928).
[110] C. H. Moore, Jr., Trans. Am. Inst. Mining Met. Petrol. Engrs. **184**, 194 (1949).
[111] A. Eucken and A. Büchner, Z. physik. Chem. **27B**, 321 (1934).
[112] C. Schusterius, Z. tech. Phys. **16**, 640 (1935).
[113] H. Büttner and J. Engl. Z. tech. Phys. **18**, 113 (1937).
[114] L. J. Berberich and M. E. Bell, J. Appl. Phys. **11**, 681 (1940).
[115] von Hippel, Breckenridge, Chesley, and Tisza, Ind. Chem. **38**, 1097 (1946).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

with frequency until the region of reststrahl vibrations is reached (wavelengths shorter than about 300 μ).

Thus the measurements of Schmidt[116,117] on single crystals, at a frequency of several hundred megacycles per second are the same as the low-frequency values. This worker obtained a value of 173 for the c direction and 89 for the a direction. These results are in agreement with the values calculated from the reflection measurements of Liebisch and Rubens.[118] The value for densely fired ceramics[115] is approximately 100. If care is taken to prevent the loss of oxygen during firing of the ceramic, the dielectric loss is very low; $\tan\delta = 0.0003$ at 1000 cy/sec and 24°C.

At a frequency of 1000 kcps, the measurements of Bunting *et al.*[119] yielded a value of 106 for the dielectric constant of pure rutile ceramic at −60°C and a value of 93 at 85°C. From these results one obtains a value of -11×10^{-4} for the temperature coefficient of the dielectric constant at 85°C and a value of -7.5×10^{-4} at −60°C.

11.2. Static Dielectric Constant and Polarizability

It is usually assumed that the local field F, effective in polarizing ions in certain types of *diagonal cubic* crystals, is given by the relation

$$F_i = E + \gamma_i P/\epsilon_0 \quad \text{(in mks units)}. \tag{11-1}$$

Following Lorentz, γ_i is usually taken to be $\frac{1}{3}$. The static dielectric constant K, and the polarization P, are defined in terms of the electric field E, and the electric displacement D, as follows:

$$D = K\epsilon_0 E = \epsilon_0 E + P. \tag{11-2}$$

The symbols N, V, n_i, F_i, P_i, and α_i are defined as follows. N = the number of unit cells per cubic meter, $V = 1/N$ = the volume of unit cell in m^3, n_i = the number of ions of type i per unit cell, F_i = the local field at ions of type i, P_i = the contribution of ions of type i to the total polarization, and α_i = the polarizability of ions of type i, and is defined by the following equation:

$$P_i = (Nn_i)\alpha_i(\epsilon_0 F_i) \tag{11-3}$$

$$n_i\alpha_i = VP_i/\epsilon_0 F_i. \tag{11-4}$$

The success of the Clausius-Mosotti relation

$$\alpha = \sum n_i\alpha_i \tag{11-5}$$

$$= 3V(K-1)/(K+2) \tag{11-6}$$

is usually taken as evidence for the validity of the Lorentz equation. To the extent that the assumptions used in deriving these results are justified, it may be shown that the polarization P in crystals having "diagonal cubic" symmetry may be expressed in terms of the sum of the polarizabilities; thus

$$P = \sum P_i = \epsilon_0 \sum F_i\alpha_i n_i/V. \tag{11-7}$$

Assuming the same F for all ions, then

$$P = (\epsilon_0 F_i/V)\sum n_i\alpha_i = (\epsilon_0 F_i\alpha/V). \tag{11-8}$$

Making use of the additivity of α implied in Eq. (11-8) Roberts[120] has used the experimental values of α of a number of ionic crystals to prepare a table of the polarizabilities of their constituent ions. The polarizability of oxygen was arbitrarily taken to be 30 A^3, so that the resulting polarizabilities of the halide ions would be nearly proportional to their volumes.

No attempt was made to separate the total polarizabilities into their electronic and ionic components. Nevertheless, it is possible to predict within a few percent the polarizabilities of a large number of ionic cubic crystals using Robert's values of α_i.

(*a*) *Temperature dependence.*—Differentiation of Eq. (11-6) with respect to temperature leads to

$$(1/\alpha)\cdot(d\alpha/dT) = 3\cdot(1/3V)dV/dT + (3K/(K+2)(K-1))\cdot(1/K)dK/dT. \tag{11-9}$$

Thus the temperature coefficient of polarizability is the sum of the temperature coefficient of volume expansion and a term proportional to the temperature coefficient of the dielectric constant. Roberts[120] quotes the values $(1/K)dK/dT = -70\times10^{-5}$, $(1/3V)dV/dT = 0.91\times10^{-5}$ and $K = 114$ at 300°K for rutile ceramic. Substitution of these values in Eq. (11-9) yields the calculated value of $(1/\alpha)d\alpha/dT = 0.90\times10^{-5}$. This result is smaller by a factor of about 30 than the corresponding value for the alkali halides. To the extent that the Clausius-Mosotti equation can be said to apply to rutile, the large negative temperature coefficient of the dielectric constant results from the thermal expansion of rutile and the small temperature coefficient of polarizability.

(*b*) *Defect structure.*—The frequency dependence of the dielectric constant and loss angle is used to investigate the defect structure of dielectric materials. Thus Srivastava[121] has studied the defect structure in slightly impure rutile crystals. The defect structure was enhanced or modified by quenching of the crystals in liquid nitrogen after heating at 820°C for three hours in dry oxygen. The resulting absorption centers were quite anisotropic, and apparently could be observed only with the electric field in the c direction.

Below 10 kc/sec, both the real part of the dielectric constant, K', and the tangent of the loss angle, $\tan\delta$, appeared to be increasing indefinitely with decreasing

[116] W. Schmidt, Ann. Phys. 9, 919 (1902).
[117] W. Schmidt, Ann. Phys. 11, 114 (1903).
[118] T. Liebisch and H. Rubens, Preuss. Akad. Wiss. Ber. 8, 211 (1921).
[119] Bunting, Shelton, and Creamer, J. Research Natl. Bur. Standards 38, 337 (1947).
[120] Shepard Roberts, Phys. Rev. 76, 1215 (1949).
[121] K. G. Srivastava, Progr. Rept. No. XXI (June,1957), and Progr. Rept. No. XXII (December, 1957), Laboratory for Insulation Research, Massachusetts Institute of Technology.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

frequency, and exceeded the values $K'=700$ and $\tan\delta=0.7$ at 100 cy/sec. Above about 1 mc/sec, K' appeared to be asymptotic to the value for the defect-free material, and $\tan\delta$ fell off to low values.

Relaxation effects occurred in the frequency range 10 kc/sec to several megacycles a second. The frequency f_{max} at which $\tan\delta$ reached a maximum value (of approximately 0.75) increased from about 8000 cy/sec at 27°C, to about 800 kcps at 278°C. These effects could be annealed out by heating in atmospheric air at moderate temperatures (about 150°C). Measurements made below about 40°C indicated that f_{max} was an exponential function of reciprocal temperature, and the data were used to yield an estimated value of 0.18 ev for the activation energy involved.

These effects were associated with the presence of traces of silver and copper (about 0.01%), and were absent in crystals free from these impurities.

Similar phenomena have been observed by Freymann,[122] who ascribed the effect to "reorientation of positive and negative lattice defects."

(*c*) *Effect of other impurities.*—Skanavi[123,124] has found that small additions of Group II oxides to rutile cause a large increase in the dielectric constant, of the order of 3×10^4. Zerfoss and co-workers[125] believed that this might be due to the formation of an "irregular sandwich of a conducting surface region and a dielectric interior with indefinite boundaries," which would produce a large space charge polarization.

The phenomenon may be explained by the overlapping of wave functions of the more closely spaced of the impurity sites. Regions of closer than average impurity separation, would be expected to occur at random. In these regions there would be produced chains of atoms which would produce high conductivity paths of limited length. Such a mechanism would constitute a slight extension of the hypothesis of Zerfoss.

Nicolini[126] has prepared a form of ceramic "rutile" with unusual dielectric properties. The material was prepared by heating "pure' titanium dioxide to 1400°C. The dielectric constant, which has a value exceeding 10 000 at 10 cy/sec, drops to its normal value of about 100 at approximately 10^5 cy/sec. $\tan\delta$ has a maximum value of 3.2 at 20 000 cy/sec.

11.3. Optical Absorption

At 4°K, "pure" rutile is transparent in the visible region at wavelengths longer than the sharp absorption edge at 4100 A. (See Figs. 10 and 11.) This edge becomes less sharp at higher temperatures, and extends to slightly longer wavelengths. The absorption coefficient rises rapidly at wavelengths longer than 5 μ.[127,128]

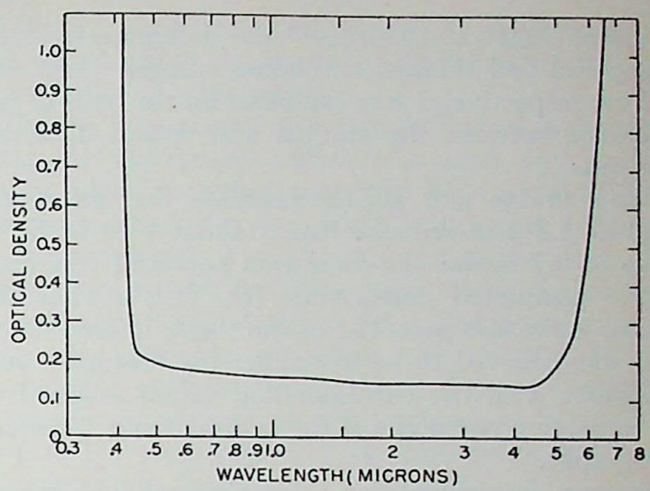


FIG. 10. Optical density of clear rutile (after Breckenridge and Hosler[138]).

After heating in hydrogen for 2.5 min at 600°C there is an increase in optical density, with a very broad maximum centered at 1.2 μ.[128]

In the far infrared, Parodi[129] found transmission lines at 30.5, 41, and 50 μ.

Breckenridge and Hosler[128] attempt to interpret the 4100-A optical edge (3.02 ev) in terms of the transition of an electron from a helium-like, neutral oxygen vacancy $(O_v\cdot 2Ti^{+3})^0$ or from a $(O_v\cdot Ti^{+3})^+$ center to the Ti^{+4} conduction band. Their assignment of the energy levels is based on evidence obtained from thermodynamics and other grounds. At higher temperatures some of these vacancies and their associated electrons are in excited states. As a result, the activation energy is supposed to be reduced and rendered less sharp.

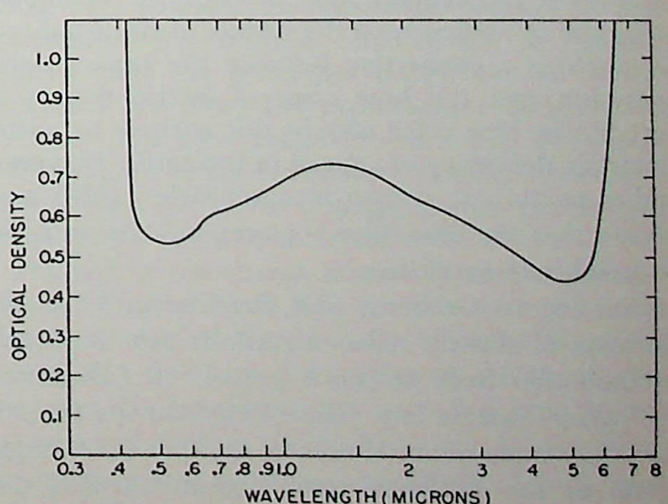


FIG. 11. Optical density of slightly reduced rutile (after Breckenridge and Hosler[138]).

[122] M. Freymann and R. Freymann, Compt. rend. **235**, 1125 (1952).
[123] G. I. Skanavi, J. Exptl. Theoret. Phys. U.S.S.R. **17**, 399 (1947).
[124] G. I. Skanavi, Doklady Akad. Nauk, S.S.S.R. **59**, 231 (1948).
[125] Zerfoss, Stokes, and Moore, J. Chem. Phys. **16**, 1166 (1948).
[126] L. Nicolini, Nature **170**, 938 (1952).
[127] D. C. Cronemeyer, Phys. Rev. **87**, 876 (1952).
[128] R. G. Breckenridge and W. R. Hosler, Phys. Rev. **91**, 793 (1953).
[129] M. Parodi, Compt. rend. **205**, 906 (1937).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

Following Mott and Gurney[130] the difference between the optical and thermal activation energies (3.02 and 0.62 ev, respectively) was explained on the basis of the difference between the statical and optical dielectric constants.

Breckenridge and Hosler associate the absorption band at 1.2 μ in reduced rutile (about 1 ev with the value of 0.2 ev for the ionization potential (thermal) of the postulated donor sites $(O_v \cdot Ti^{+3})^+$. Thus the optical ionization potential of the singly ionized donor sites was believed to be five times the thermal value, consistent with the corresponding values assigned to the ionization potentials of the neutral donor (previous paragraph).

Support is claimed for this assignment of energy levels by experiments on the bleaching of colored rutile on cooling to liquid air temperatures mentioned briefly in the same paper.

11.4 Color

Weyl and Forland[131] have discussed (i) the color changes introduced by traces of ions such as Nb^{+5}, Ta^{+5}, Sb^{+5}, W^{+6}, which are usually considered colorless, and by very small traces of Fe^{+3} in rutile; (ii) the phototropy of rutile (reversible darkening when exposed to light) which may be obtained by incorporating certain impurities into rutile, as for example, a combination of the oxides of iron and niobium; and (iii) the oxidizing effect which titanium dioxide may exert on other materials when it is irradiated by ultraviolet light.

The presence of the Fe^{+3} ion introduces a rose color, that of the Nb^{+5} ion a blue-gray color into rutile. The blue-gray color is also present in reduced rutile, but unlike the effect of the Nb^{+5}, the discoloration of reduced rutile may be bleached by reoxidation.

Sandler's experiments on the rate of *ortho-para* conversion of hydrogen on the surface of rutile indicate that there is a connection between the surface paramagnetism and the blue color of lattice defects in rutile.[132] The blue color may be due entirely to a stoichiometric deficiency of oxygen in the rutile. However, Sandler points out certain inconsistencies which may indicate that the blue color is associated also with the the presence of impurities.

According to Gebhardt and Herrington,[133] the discoloration of slightly reduced rutile is due to organic contamination from stopcock grease, etc. Degassing of a "clean" sample in a "clean" vacuum system produced no discoloration. However, if either the vacuum system or the specimen were contaminated, a discoloration resulted. At temperatures of 300 to 550°C the loss of oxygen was estimated to be about 10^{-8} mole/g of TiO_2.

When heated in air to high temperatures, powdered rutile becomes lemon-yellow but returns to its original whiteness on cooling again to room temperature.[134] At 900°C there is a maximum in the emission spectrum at about 4700 A.

11.5 Refractive Indices

Measurements of the index of refraction (normal ray) in the visible region, were made by Bärwald[135] and Schroeder.[136] Cronemeyer[127] has obtained good agreement with the earlier work, and has given the result of his measurements together with values calculated from reflection measurements (Liebisch and Rubens[118]) and from transmission measurements.

Using prisms cut from single crystals, DeVore[137] has measured the refractive indices of rutile in the wavelength range 4250 to 15 000 A. DeVore has expressed his results by means of the formulas

$$n^2 = 7.197 + 3.322 \times 10^7/(\lambda^2 - 0.843 \times 10^7)$$

for the extraordinary ray and

$$n^2 = 5.913 + 2.441 \times 10^7/(\lambda^2 - 0.803 \times 10^7)$$

for the ordinary ray. The results are said to be in good agreement with those of Bärwald.[135]

Radhakrishnan[138] has used the method of Bragg to calculate the refractive indices of rutile. This worker has also used the measurements of Schroeder[136] to derive the values of the constants a, b, c, and λ_0 in the dispersion formula for rutile

$$n^2 - 1 = a - c\lambda_0^2 - b\lambda^2/(\lambda^2 - \lambda_0^2)$$

in which n is the index of refraction at wavelength λ. Values calculated using this relation agreed with observation within the experimental error.

11.6 Reflection Coefficient

The reflection coefficient of natural rutile crystals has been measured by Liebisch and Rubens[118] in the range 1.4 to 300 μ. Their results appear in Fig. 12.

12. ADSORPTION AND SURFACE LAYERS

In this section no reference is made to the extensive literature concerning adsorption of various gases and vapors on rutile powder, with the exception of certain experiments which yield direct information concerning the defect structure of the surface of rutile.

From the measured rate of *ortho-para* conversion of hydrogen at the surface of rutile (discussed also in

[130] N. F. Mott and R. W. Gurney, *Electronic Processes in Ionic Crystals* (Oxford University Press, New York, 1948).
[131] W. A. Weyl and Tormod Forland, Ind. Chem. 42, 257 (1950).
[132] Y. L. Sandler, J. Phys. Chem. 58, 54 (1954).
[133] J. Gebhardt and K. Herrington, J. Phys. Chem. 62, 120 (1958).
[134] G. I. Sinyapkina, Zhur. Eksptl. i Teoret. Fiz. 19, 581 (1949).
[135] W. Bärwald, Z. Krist. 7, 168 (1883).
[136] A. Schroeder, Z. Krist. 67, 485 (1928).
[137] J. R. DeVore, J. Opt. Soc. Am. 41, 416 (1951).
[138] R. S. Radhakrishnan, Proc. Indian Acad. Sci. A35, 117 (1952).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by eGangotri Foundation, USA

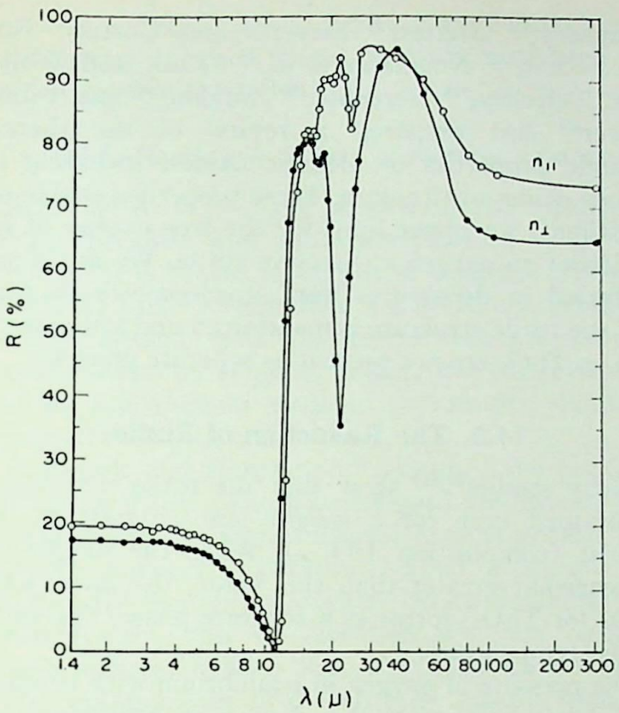


FIG. 12. Reflection coefficient in percent for natural single crystals of rutile (after Liebisch and Rubens[118]).

Sec. 11.4), Sandler[139] has obtained information concerning the state of adsorbed oxygen. The oxygen adsorbed on rutile at 90°K is apparently in molecular form, and the correspondingly high conversion rate is ascribed to the paramagnetism of this form of oxygen. On heating the specimen to room temperature, dissociation of some of the oxygen evidently occurred, leading to re-oxidation of the surface and a disappearance of the blue color. Reyerson and Honig[140] have investigated the diamagnetic susceptibility of NO_2-N_2O_4 adsorbed on rutile. To obtain a constant value for the gram susceptibility of the absorbate these workers found it necessary to assume that the (positive) susceptibility of the rutile decreased during adsorption. Thus it seemed reasonable to suppose that the rutile was initially partially reduced, but became re-oxidized by the NO_2-N_2O_4. The true value (after re-oxidation) of the susceptibility of rutile estimated in this way was 0.061×10^{-6} cgs units per gram. This interpretation is strengthened by the observed permanent increase in weight of the specimen, and by spectroscopic analysis of the equilibrium gases.

12.1. Photoconductivity in Surface Layers

The curves of spectral distribution of photoconductivity of most semiconductors exhibit a sharp rise in the vicinity of the absorption edge. As the wavelength of the incident light decreases, the response reaches a maximum value, and begins to decrease while the absorption coefficient is still rising. At small intensities the photocurrent is usually proportional to the light intensity at all wavelengths.

On the long wavelength side of the response maximum, the absorption coefficient is small, and much of the incident radiation traverses the crystal without being absorbed. On the short wavelength side, the penetration depth is small, and most of the absorption takes place in the surface region, where the electron recombination velocity†† is high. The wavelength of maximum response is determined by the relative importance of these mechanisms. Thus the response maximum will depend on the thickness of the crystal, and on the temperature, if the absorption coefficient is temperature-dependent.

Cronemeyer[127] has measured the photocurrent of thin crystals of rutile as a function of the wavelength of the incident radiation. The response increased rapidly with the energy of the incident photons in the region of 2.9 ev (4300 A). The edge of the photoconductivity curve is shifted to slightly lower energies as the temperature is increased from 103 to 295°K. Because of the high absorption coefficient above 3 ev (4100 A), and because of the important role of surface effect in this region, the interpretation of the results is difficult for higher photon energies.

At 1.2 μ, the response is smaller than the maximum response by a factor of 10^6.

13. BAND STRUCTURE OF RUTILE—CONCLUSIONS

Estimates of the width of the energy gap from measurements of electrical conductivity in the intrinsic range involve an assumption that the density-of-states effective masses of the valence and conduction bands are identical.

The measurements of electrical conductivity made in the high temperature range (700 to 1100°C) where presumably the intrinsic conductivity is measured, indicate that the forbidden energy gap of rutile is between 3 and 4 ev. The experimental value obtained by Earle (3.4 ev), Hauffe (3.9 ev), and Cronemeyer (3.06 ev) are discussed in another section. Cronemeyer's investigation of optical absorption and photoconductivity yield a value of about 2.8 to 3.1 ev.

The sign of the charge carriers in semiconductors is usually obtained from Hall or thermopower measurements. Because of the difficulty of measuring the potential at a point on the surface of a high resistance specimen, Hall measurements have not been made on specimens of rutile in the intrinsic range. Measurements of thermopower on rutile specimens of high purity indicate that the charge carriers are probably electrons.

The measurements of Breckenridge and Hosler

[139] Y. L. Sandler, J. Phys. Chem. 58, 54 (1954).
[140] L. H. Reyerson and J. M. Honig, J. Am. Chem. Soc. 75, 3920 (1953).

†† The surface recombination velocity is defined as "the rate at which electrons disappear from the interior across the surface, divided by the density of electrons in excess of equilibrium in the interior."

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

indicate that the mobility of reduced specimens is about 0.1 cm^2 $volt^{-1}$ sec^{-1} at 500°K. The value for intrinsic rutile may be approximately the same. By combining this value of mobility with the electrical conductivity measurements of Cronemeyer, one obtains an effective electron mass ratio of the order of 1000. However, values of the electron concentration evaluated from the combined Hall and conductivity data of Breckenridge and Hosler (see Sec. 8) yield effective mass ratios of the order of unity.

An attempt has also been made to discuss the band structure of reduced rutile with the inclusion of a donor band. The measurements of the magnetic susceptibility of rutile containing a stoichiometric deficiency of oxygen made by Ehrlich could be explained by the presence of such a donor band approximately 0.04 ev in width. Assuming the existence of this band (in which the mobility is expected to be low, and the effective electron mass high), then the donor band is probably 0.05 to 0.20 ev below a "conduction band" (in which the effective mass is considerably lower). The values of the energy separation of these bands calculated from the conductivity, Hall effect, and thermopower measurements discussed in earlier sections are given after the author's names. Electrical conductivity of reduced rutile: Kataoka and Suzuki (0.12 to 0.15 ev). Thermopower: Kataoka and Suzuki (0 to 0.13 ev). Hall data: Breckenridge and Hosler (0.025 to 0.05 ev). Optical absorption: An absorption band characteristic of reduced rutile was observed by Cronemeyer in the range 1 to 3 μ (0.4 to 1.2 ev). In order to compare these energies with the other data, they may have to be reduced by a factor of about 5.1 to allow for the difference between the optical and statical dielectric constants, as discussed in Sec. 11.3. This correction would place the donor band about 0.08 to 0.24 ev below the conduction band.

The evidence for the presence of a donor band is not conclusive. Consequently, the data have also been discussed below in terms of noninteracting sites. On the basis of the simple model and the assumption of noninteracting donor sites, the electrical conductivity measurements lead to the values of donor ionization potentials: Kataoka and Suzuki (0.29 to 0.35 ev), Boltaks (less than 0.2 ev). Thermopower: Kataoka and Suzuki (0.02 to 0.13 ev); Henisch (0.09 ev). Hall measurements: Breckenridge and Hosler (0.10 to 0.15 ev).

14. THERMODYNAMIC PROPERTIES AND REDUCTION

14.1. Thermodynamic Properties

The thermodynamic properties, free energy of formation, etc., of the various oxides (and many other compounds of titanium) are fully listed in the publication: The Properties of Titanium Compounds and Related Substances, ONR Report ACR-17 (October, 1956). These tables were prepared from the results of Humphrey,[141] Mixter,[142] Sieverts and Gotta,[143] Roth and Becker,[144] Neumann *et al.*,[145] Roth and Wolf,[146] Foëx,[147] Kelley,[148] Shomate,[149] Naylor,[150] and others. Brewer[151] has prepared a review of the thermodynamic properties of metallic oxides, including the various oxides of titanium. These properties enable one to estimate an upper limit for the free energy of formation of an oxygen vacancy in rutile. We are chiefly interested in departures from stoichiometry so small that the rutile structure is maintained and lower oxides such as Ti_3O_5 are not present as separate phases.

14.2. The Reduction of Rutile

X-ray studies[80,152] show that the rutile structure is maintained even for a weight loss as great as two percent (composition $TiO_{1.90}$). When the weight loss is somewhat greater than this value, the lower oxide Ti_3O_5 (or Ti_2O_3) forms as a separate phase, depending on the temperature.

The pressure of oxygen in equilibrium with titanium oxide decreases with the degree of reduction, the decrease being most rapid when the composition is close to the relative oxygen content represented by the formulas Ti_2O_3, Ti_3O_5, and TiO_2. For values of weight loss up to two percent, the degree of reduction may apparently be controlled[163] by heating the specimen of rutile in oxygen at reduced pressure, in a vacuum system free from contamination by stopcock grease. For high degrees of reduction, the equilibrium oxygen pressure is inconveniently low for the use of this simple method, and it becomes necessary to reduce the specimen by chemical means.

The oxygen pressure may be established at the desired low value by heating the specimen in a mixture of hydrogen and water vapor of known composition. The equilibrium constant of this reaction is known with accuracy,[153] and the equilibrium oxygen pressure may be calculated from the following equation:

$$\log K_p = \log p_{O_2} (p_{H_2}/p_{H_2O})^2 = 5.9 - 25\,900/T.$$

Alternatively, the CO-CO_2 reaction may be used to establish the desired oxygen pressure around the specimen.[152] A third method involves the heating of the titanium oxide in a sealed container with some

[141] G. L. Humphrey, J. Am. Chem. Soc. **73**, 1587 (1951).
[142] W. G. Mixter, Am. J. Sci. **27**, 393 (1909).
[143] A. Sieverts and A. Gotta, Z. anorg. u. allgem. Chem. **199**, 384 (1931).
[144] W. A. Roth and G. Becker, Z. physik. Chem. **A159**, 1 (1932).
[145] Neumann, Kröger, and Kunz, Z. anorg. u. allgem. Chem. **218**, 379 (1934).
[146] W. A. Roth and U. Wolf, Rec. trav. chim. **59**, 511 (1940).
[147] M. Foëx, Bull. soc. chim. **11**, 6 (1944).
[148] K. K. Kelley, U. S. Bur. Mines Bull. 383 (1935).
[149] C. H. Shomate, J. Am. Chem. Soc. **69**, 218 (1947).
[150] B. F. Naylor, J. Am. Chem. Soc. **68**, 1077 (1946).
[151] L. Brewer, Chem. Revs. **52**, 1 (1953).
[152] Assayag, Dode, and Faivre, Compt. rend. **240**, 1212 (1955).
[153] F. D. Richardson and J. H. E. Jeffes, J. Iron Steel Inst. (London) **160**, 261 (1948).

CC-0. Panjab University Chandigarh Collection. Digitized by eGangotri Foundation, USA

pieces of a metal which forms a very stable oxide. Experiments in which these methods have been used are discussed in the following.

14.2.1. Heating in Oxygen at Reduced Pressure

Assayag *et al.*[152] have measured the equilibrium oxygen pressure and the corresponding weight loss when rutile is heated in oxygen to temperatures above 1000°C.‡‡ Little information was given concerning the vacuum system used. Because of the possibility of contamination by stopcock grease, and the presence of CO in the vacuum system, the results should be accepted with caution.

Gebhardt and Herrington[154] obtained the usual discoloration when rutile was degassed for sixteen hours at 400 to 600°C in a conventional vacuum system, using greased stopcocks. However, when a sample was cleaned by heating in oxygen before degassing in a clean system, this discoloration was not obtained. These workers believe that the discoloration results from contamination of the rutile by organic impurities originating in the stopcock grease. The "clean" vacuum system made use of a mercury diffusion pump and mercury cutoffs instead of stopcocks. The conclusions were substantiated by mass spectrometric analysis and the detection of CO and CO_2 in the vacuum system.

14.2.2. Use of the CO-CO_2 *Equilibrium*

Assayag *et al.*[152] have used the equilibrium between CO and CO_2 to control the degree of reduction of rutile corresponding to compositions in the range $TiO_{1.908}$ to $TiO_{1.977}$ (i.e., 0.4% to 2% loss in weight). They have also shown by x-ray methods that this material retained the rutile structure.

14.2.3. Use of the Hydrogen-Water Vapor Equilibrium

Nasu[155–157] has used the hydrogen-water vapor equilibrium in the reduction of rutile in the temperature range 1022 to 1282°K. A specimen of powdered rutile was placed in a platinum crucible and suspended in an iron reaction chamber at the end of a quartz spring. The approach to equilibrium could be determined by the change in weight. A mixture of hydrogen and water vapor of unstated composition was introduced to the specimen. Later the composition of a sample of the equilibrium mixture was determined by measuring its pressure with a McLeod gauge both before and after the removal of the water vapor by condensation with liquid nitrogen, and the ratio of the partial pressures in the equilibrium mixture was then calculated.

Nasu[155–157] found that the equilibrium oxygen pressure remained constant from one percent to about 9.7% reduction, at which point its value fell abruptly. The absence of variation in the equilibrium constant in this range of reduction seems to indicate that the lower oxide existed as a separate phase, and the value of 9.7% corresponds closely to the loss in weight corresponding to the formation of Ti_2O_3. These results seem to set an upper limit of one percent to the weight loss which rutile can sustain at these temperatures without leading to the presence of aggregates of a different phase, i.e., when speaking of slightly reduced rutile, a weight loss of less than one percent is implied.

Unfortunately, the equilibrium ratios calculated from more recent calorimetric and free energy data are considerably higher than Nasu's values. Moreover, as Kelley has pointed out, Nasu's high-temperature data correspond exactly to the p_{H_2O}/P_{H_2} ratio for the FeO-Fe system, which fact may be related to Nasu's use of an iron reaction chamber.

14.2.4. Use of the Alkaline Earth Metals

Kubaschewski and Dench[158] have investigated the partial free energy of the titanium-oxygen system at 1000 and 1200°C between the limits of zero and 38% oxygen by weight. A quantity of barium, calcium, or magnesium was mixed with a quantity of one or other of the oxides of titanium. The mixture was sealed in a steel container and fired at the desired temperature. After cooling, the alkaline earth oxide was dissolved with hydrochloric acid and the remaining powder was analyzed for O; N, H, Fe, Ba, Ca, or Mg, and Ti. The equilibrium oxygen pressure was assumed to be that of the alkaline earth oxide at the temperature of the reaction. From these data a plot was made of partial free energy against composition of the oxide. In this paper, the effect of the formation of alkaline earth titanate was not discussed. The pressure of oxygen in equilibrium with a quantity of alkaline earth titanate is probably lower than that in equilibrium with the alkaline earth oxide.

14.2.5. Other Methods

Ehrlich[80] prepared specimens of TiO_x with x greater than 1.97 by heating appropriate mixtures of rutile and titanium metal or TiO to 1500–1600°C. Wyss[159] investigated the reduction of rutile by hydrogen and reported that at 1575°K, Ti_3O_5 is formed. The absence of further reduction under these conditions seems to indicate, as pointed out by Brewer,[151] that Ti_2O_3 is

‡‡ In the temperature range $T=1318$ to 1531°K, and $p=10^{-2}$ to 10^{-4} atmos, the results of these workers may be expressed very roughly by the following expression for the fractional weight loss f.

$$\log f = -\tfrac{1}{3}\log p_{O_2} - 5000/T - 0.10.$$

[154] J. Gebhardt and K. Herrington, J. Phys. Chem. **62**, 120 (1958).
[155] Nobuyuki Nasu, Sci. Repts. Tôhoku Univ. **12**, 371 (1935).
[156] Nobuyuki Nasu, J. Chem. Soc. (Japan) **56**, 659 (1935).
[157] Nobuyuki Nasu, Sci. Repts. Tôhoku Univ. **XXV**, 510 (1936).
[158] Kubaschewski and Dench, J. Inst. Metals **82**, 87 (1953).
[159] R. Wyss, Ann. chim. **3**, 215 (1948).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

more stable relative to Ti_3O_5 than Humphrey's data appear to indicate.[141]

Kelley has investigated the reduction of rutile in an atmosphere of hydrogen-water vapor, but his results have not been published. As reported by Brewer,[151] the equilibrium pressure ratio p_{H_2O}/p_{H_2} changes abruptly at the composition Ti_3O_5. On the other hand, Welch[160] found no evidence for Ti_3O_5 when rutile is reduced by hydrogen at 1400°K. At 1700°K, Ti_3O_5 can be formed, in confirmation of Humphrey's conclusion that Ti_3O_5 is stable only at high temperatures.

Russian workers have investigated the reduction of metal oxides by hydrogen, and by alcohols.[161] Freundlich and Bichara[162] have used CaH_2 to reduce rutile. In his paper "Reduction of metallic oxides by carbon in boric oxide and borax melts," Khundkar[163] reports that rutile is not reduced by active carbon at 1000°C.

[160] A. J. E. Welch, 120th Meeting of the American Chemical Society, New York (September, 1951).

[161] V. A. Komarov, Uchenye Zapiski Leningrad Univ. No. 169, Ser. Khim. Nauk No. 13, 29–35 (1953).

[162] W. Freundlich and M. Bichara, Compt. rend. **238**, 1324 (1954).

ACKNOWLEDGMENTS

The writer is grateful to A. D. Franklin, H. P. R. Frederikse, H. S. Peiser, and J. B. Wachtman, Jr., of the National Bureau of Standards, who have read the manuscript, and have made numerous valuable suggestions.

Acknowledgment, with thanks, is extended to the following authors and journals for permission to use their diagrams: L. Pauling and the Cornell University Press for Fig. 2. W. H. Baur and Acta Crystallographic for Fig. 3. H. P. Hanson, J. R. Knight, and the Physical Review, for Fig. 4. R. G. Breckenridge, W. R. Hosler, and the Physical Review, for Figs. 1, 8, 9, 10, and 11. T. Liebisch, H. Rubens, and die Deutsche Akademie der Wissenschaften for Fig. 12.

[163] M. H. Khundkar, J. Indian Chem. Soc. **29**, 467 (1952).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

Comparison of the Debye θ Determined from Elastic Constants and Calorimetry

G. A. Alers and J. R. Neighbours

Scientific Laboratory, Ford Motor Company, Dearborn, Michigan

INTRODUCTION

As the temperature of a solid approaches 0°K the simple Debye continuum model should be applicable and the lattice specific heat calculated from the elastic constants using this model should agree with that measured by calorimetric techniques. Now that accurate low-temperature specific heat and elastic constant data are available, it is possible to compare critically the values of the Debye characteristic temperature θ, which these two methods yield. Usually, the calorimetrically determined θ is obtained from the slope of the straight line in a plot of C/T *versus* T^2 where C is the measured specific heat and T the absolute temperature. Such a plot separates the contribution of the electrons and is a straight line only if the lattice specific heat varies with T^3 as predicted by the Debye model. To determine θ from the elastic constants requires the calculation of the three sound wave velocities as a function of direction in the crystal lattice and then the averaging of the reciprocal of each velocity cubed over all directions. It is possible to do this by hand for hexagonal materials, but cubic materials require a long and elaborate calculation which is reasonable only if an electronic computer is available. Without a computer the tables prepared by de Launay[1] or the approximate methods developed by Quimby and Sutton,[2] and by Betts *et al.*[3] may be used.

Various models have been proposed to determine the lattice specific heat in the temperature range where the Debye model is expected to become inapplicable. These models consider a discrete lattice rather than a continuum and use measured elastic constants to define the effective force constants which act upon any given lattice point. With these force constants, the dynamics of the entire lattice can be worked out and the specific heat calculated.

This paper evaluates the different methods of calculating θ from the elastic constants and compares the resulting values with the results of calorimetric measurements.

DETERMINATION OF θ_0

The necessary data for calculating the value of Debye θ at 0°K, hereafter denoted by θ_0, are shown in Table I. In this table, the low-temperature elastic constants of copper,[4] silver,[5] gold,[5] zinc,[6] magnesium,[7] and lithium fluoride[8] are extrapolated from measurements performed at 4.2°K while those of aluminum,[9] lithium,[10] sodium chloride,[11] and potassium bromide[12] are extrapolated from measurements near 77°K. The 0°K atomic volumes and densities are calculated from the lattice constants appearing in the ASTM card catalog and the following thermal expansion measurements: Nix and McNair[13] for the noble metals and aluminum, Pearson[14] for lithium, Grüneisen and Goens[15] for zinc, and Henglein[16] for potassium bromide and sodium chloride. The data for magnesium and lithium fluoride are taken from the papers cited.

Table II lists values of θ_0 as calculated from the data in Table I by various methods along with the results of calorimetric measurements. The first column shows the θ_0 obtained from de Launay's tables[1] using case 2 and the Sterling central difference interpolation formula[17] in two dimensions. A graphical interpolation may cause the results to differ by about 0.3%. These tables are

Table I. Constants of certain materials extrapolated to absolute zero. The elastic constants are in units of 10^{12} dyne/cm². Ω is the atomic volume and ρ the density at 0°K.

	C_{11}	C_{12}	C_{44}	C_{13}	C_{33}	Ω^{-1} (10^{23} cm^{-3})	ρ (g cm^{-3})
Elastic constants extrapolated from 4.2°K							
Copper	1.7620	1.2494	0.8177			8.553	9.018
Silver	1.3149	0.9733	0.5109			5.938	10.635
Gold	2.0163	1.6967	0.4544			5.953	19.488
LiF	1.246	0.424	0.649			12.292	2.646
Magnesium	0.6348	0.2594	0.1842	0.2170	0.6645	4.407	1.779
Zinc	1.7909	0.375	0.4595	0.5537	0.6880	6.708	7.279
	1.7696	0.3480	0.4589	0.528	0.6848		
Elastic constants extrapolated from the 77°K region							
Aluminum	1.230	0.708	0.3090			6.109	2.734
Lithium	0.1574	0.1333	0.1158			4.748	0.5471
NaCl	0.5750	0.0986	0.1327			4.619	2.241
KBr	0.418	0.056	0.052			2.854	2.819

[1] J. de Launay, *Solid State Physics* (Academic Press, Inc., New York, 1956), Vol. 2, p. 286.
[2] S. L. Quimby and P. M. Sutton, Phys. Rev. **91**, 1122 (1953), P. M. Sutton, Phys. Rev. **99**, 1826 (1955).
[3] Betts, Bhatia, and Wyman, Phys. Rev. **104**, 37 (1956).
[4] W. C. Overton and J. Gaffney, Phys. Rev. **98**, 969 (1955).
[5] J. R. Neighbours and G. A. Alers, Phys. Rev. **111**, 707 (1958).
[6] The first set of constants are from G. A. Alers and J. R. Neighbours, J. Phys. Chem. Solids, **7**, 58 (1958). The second set are the recent measurements by C. W. Garland and R. Dalven, Phys. Rev. **111**, 1232 (1958).
[7] L. J. Slutsky and C. W. Garland, Phys. Rev. **107**, 972 (1957).
[8] C. V. Briscoe and C. F. Squire, Phys. Rev. **106**, 1175 (1957).
[9] P. M. Sutton, Phys. Rev. **91**, 816 (1953).
[10] H. C. Nash and C. S. Smith, Bull. Am. Phys. Soc. Ser. II, **3**, 123 (1958).
[11] W. C. Overton and R. T. Swim, Phys. Rev. **84**, 758 (1951).
[12] J. K. Galt, Phys. Rev. **73**, 1460 (1948).
[13] F. C. Nix and D. McNair, Phys. Rev. **60**, 597 (1941); Phys. Rev. **61**, 74 (1942).
[14] W. B. Pearson, Can. J. Phys. **32**, 708 (1954).
[15] E. Grüneisen and E. Goens, Z. Physik **29**, 141 (1924).
[16] A. Henglein, Z. physik. Chem. **115**, 91 (1925).
[17] J. de Launay (private communication).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

TABLE II. Debye temperature at zero degrees Kelvin.

	de Launay	Quimby and Sutton	Betts *et al.*	Numerical integration	Calorimetric	Reference
Elastic constants extrapolated from 4.2°K						
Copper	344.5	344.4	345.4	344.4	343.8 346.7 345.1	a b c
Silver	226.4	226.6	227.1	226.4	225.3 226.5 226.2	a b d
Gold	161.7	162.1	162.2	161.6	164.6 164.8 164.6	a b e
LiF	734.4	734.6	734	734.1	737 743	f g
Magnesium	···	···	385.4	385.8	406 403.3	h i
Zinc	···	···	327.1	327.1 328.8	302 309	j
Elastic constants extrapolated from 77°K						
Aluminum	427.4	428.5[k]	428.7	428.2	375 408 419	l m
Lithium	334.6		338.4	335.9	369	n
KBr	···		171.7	172.8	174	o
NaCl	···	321.6	321	321.9	320	p

[a] See reference 19.
[b] See reference 30.
[c] J. A. Rayne, Australian J. Phys. 9, 189 (1956).
[d] J. A. Rayne, Proc. Phys. Soc. (London) 69, 482 (1956).
[e] See reference 29.
[f] See reference 20.
[g] T. H. K. Barron and J. A. Morrison, Can. J. Phys. 35, 799 (1957).
[h] See reference 21.
[i] See reference 28.
[j] See reference 22.
[k] The calculated value listed in reference 9 is incorrect. The corrected value calculated by Dr. Sutton is listed above.
[l] See reference 23.
[m] See reference 24.
[n] See reference 25.
[o] See reference 26.
[p] See reference 27.

based on a computer calculation and the values calculated should be exact except for errors introduced by interpolation between the tabulated values. Since the published tables apply only to cubic materials with $(C_{11}-C_{12})<2C_{44}$ there are no entries for potassium bromide and sodium chloride nor for the hexagonal metals. The second column gives the results obtained from approximating the velocity function with a six-term power series[2] and the third column shows the results obtained if a six-term expansion in cubic harmonics[3] is used. The latter expansion technique has been applied to hexagonal materials by using spherical harmonics[18] and was used to complete column three for zinc and magnesium.

The IBM 650 digital computer of the Ford Motor Company Research Department was programed to find the average velocity function using 120 points spread uniformly over the unit triangle for cubic materials and using a one degree interval for the single angle variable in hexagonal materials. The resulting values of θ_0 are given in the column headed "Numerical integration." The two entries for zinc are calculated from the two available sets of 4.2°K elastic constants[6] in Table I. Increasing the number of points considered in the unit triangle to 933 increased the value of θ_0 for copper by 0.17°K or 0.06% and decreased the θ_0 value for lithium by 0.49°K or 0.15%. Since the values from a 120-point calculation differ so slightly from the results of the longer calculation, and since these differences are much less than the errors expected from errors in the elastic constants, it is felt that the 120-point calculation with its consequent saving in computer time is sufficiently accurate.

Table II shows that all the foregoing methods for calculating θ_0 from elastic constants are essentially in agreement. In every case the differences are less than one percent. The columns headed "de Launay" and "Numerical integration" are results of essentially the same calculation and should agree closely, as they do. The differences between these columns are indicative of the errors introduced by interpolation and by the method of numerical integration. The large difference apparent for aluminum is presumably due to our slight extrapolation of de Launay's tables to cover the particular values of elastic constants for this metal. For lithium, the large difference (1.3°K) is not due to the inadequacy of the 120-point division of the unit triangle, but to some other unknown cause.

Roughly the ease of calculation of θ_0 by any one method is in the same order as the columns in Table II. If one uses graphical interpolation, the method of de Launay is outstanding in its simplicity. Although the two series methods are about equally involved, it is our experience that the value of θ_0 resulting from the method of Betts and co-workers is apt to be more nearly mistake-free, probably because the individual steps can be checked separately.

Since some cubic materials and all hexagonal materials fall outside the range of de Launay's published tables and since the series methods are probably poorer approximations, we take the column headed "Numerical integration" to be the best values of θ_0 calculated from elastic constants.

The uncertainty of these values is determined by the absolute accuracy of the elastic constants. With the computer program setup it is a simple matter to determine how sensitive the final θ_0 value is to errors in each individual constant. Table III shows the values of θ_0 for aluminum (low anisotropy) and for gold (average

TABLE III. Changes in Debye temperature with changes in elastic constants.

	Numerical integration	1.005 C_{11}	1.005 C_{12}	1.005 C_{44}	1.01 C_{11}
Aluminum	428.2	429.2	427.7	428.8	
Gold	161.6	162.7	160.7	161.8	
Zinc	327.1				327.3

[18] See reference 3, p. 43.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

anisotropy) which are obtained by numerical integration if only one elastic constant is increased by 0.5%. A 0.5% change in elastic constants causes θ_0 to change by less than 1%. Also listed is the θ_0 value obtained for zinc when the elastic constant C_{13} is increased by 1%. This particular constant is difficult to measure accurately so it is encouraging to observe that θ_0 is not very sensitive to its uncertainties. From pulsed ultrasonic measurements, the absolute values of the measured elastic constants are estimated to be in error by about one-half percent. Since the errors in θ_0 are about the same size as the errors in the elastic constants, the calculated Debye temperature may be expected to be accurate to about one percent. This conclusion is supported by the fact that the θ_0's calculated from the two sets of low-temperature elastic constant data available on zinc differ by only 0.7%.

COMPARISON OF θ_0 WITH CALORIMETRIC RESULTS

The last column in Table II gives the values of θ_0 determined from specific heat measurements[19–27] in the 1 to 4°K temperature range. Included are several as yet unpublished results[28,29] as well as a new calculation,[30] using the 1955 temperature scale, of data already in the literature.[19] These values have been obtained by fitting the data on a C/T *versus* T^2 plot with a straight line by the method of least squares. The slope of this line defines the calorimetric θ_0. Small differences in the numerical value of the calorimetric and elastic θ_0's may arise simply because the elastic constant line is not the best least squares fit to the data. Thus the best way to compare the two results is to plot the calorimetric data and superimpose on it the line defined by the elastic constants. Since only the slope of the line is defined, it must be forced to pass through the calorimetric data at some point and disagreement is then apparent if the line and the data diverge.

Rather than plot a separate C/T *versus* T^2 graph for each material listed in Table II, the calorimetric data have been normalized to allow several sets of data to be presented in one graph. This is shown in Figs. 1 and 2 where $C\theta/T$ is plotted against $(T/\theta)^2$. On such a plot vertical lines are lines of constant T/θ. If the specific heat data are described by the sum of a linear and a cubic temperature term then, in a plot like Fig. 1, the data will define a straight line with an intercept of $\gamma\theta$ (where γ is the coefficient of the linear term) and a slope determined by the ratio of the normalizing θ to the θ which best fits the data. In constructing these figures, the values of normalizing θ were taken from the numerical integration column of Table II. The lines are drawn with the slope predicted by the Debye theory and so as to fit best the lower end of the data. We have chosen not to plot data in which there are good reasons for the large and clear differences between the calorimetric and elastic θ_0's. These are sodium chloride,[27] aluminum,[23,24] lithium,[25] and potassium bromide.[26] For sodium chloride and aluminum the specific heat data are quite scattered and can easily be described with the elastic θ_0. For lithium the lines defined by the elastic data and the calorimetric data diverge completely, but this is probably owing to the presence of a crystal structure change[31] between 77°K where the elastic constants were measured and 4°K where the specific heat was measured. In the case of potassium bromide the calorimetric measurements extend only to temperatures of the order of $\theta/50$ which is not low enough for a valid comparison.

Figure 1 shows the plot of $C\theta/T$ *versus* $(T/\theta)^2$ for the specific heat measurements of the noble metals[19] and lithium fluoride.[20] It is apparent that the elastic θ_0 defines a straight line which describes the calorimetric data for copper, silver, and lithium fluoride to within

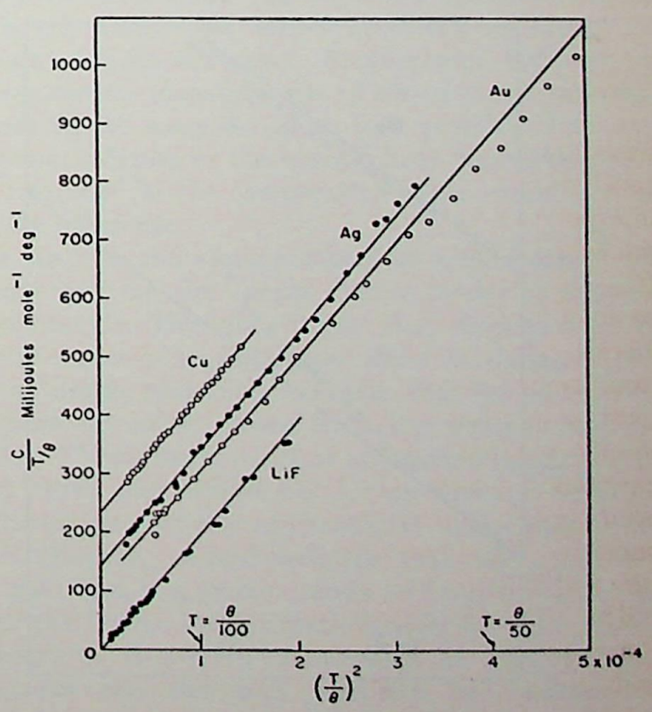


FIG. 1. Comparison of the specific heat calculated from Debye theory (solid lines) with the measured specific heats of cubic materials.

[19] Corak, Garfunkel, Satterthwaite, and Wexler, Phys. Rev. 98, 1699 (1955).
[20] D. L. Martin, Phil. Mag. 46, 751 (1955).
[21] P. L. Smith, Phil. Mag. 46, 744 (1955).
[22] G. Seidel and P. H. Keesom, Bull. Am. Phys. Soc. Ser. II, 3, 17 (1958). By assuming that θ is temperature dependent between 1 and 4°, Seidel finds that a least squares fit to his data would give $\theta_0=309$°K. We are indebted to Mr. Seidel for communicating this result to us. See also Phys. Rev. 112, 1083 (1958).
[23] Howling, Mendoza, and Zimmerman, Proc. Roy. Soc. (London) A229, 86 (1955).
[24] J. A. Kok and W. H. Keesom, Physica 4, 835 (1937).
[25] L. M. Roberts, Proc. Phys. Soc. (London) E70, 744 (1957).
[26] Barron, Berg, and Morrison, Proc. Roy. Soc. (London) 242, 472 (1957).
[27] Morrison, Patterson, and Dugsdale, Can. J. Chem. 33, 375 (1955).
[28] J. Rayne, J. Phys. Chem. Solids (to be published).
[29] J. E. Zimmerman (unpublished data, 1958).
[30] J. Skalyo and A. Arrott (to be published). These results are based on the data of reference 19 but are recalculated using the 1955 temperature scale and a least squares fit to the data.
[31] C. S. Barrett, Acta Cryst. 9, 671 (1956).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

the experimental scatter of the data. For gold, the elastic line is clearly not the best straight line fit to the data. This difference is just outside the experimental errors arising from elastic constant errors. It may be that simple Debye theory is not appropriate over the whole temperature range given in the figure in which case the elastic constant line should fit the data only below a certain temperature. From this point of view, it may be concluded that the elastic θ_0 and Debye theory successfully describe the calorimetric data below temperatures of approximately $\theta/60$.

The disagreement between calorimetric and elastic θ_0's is more explicit for the hexagonal metals since both cases examined show large deviations. Figure 2 shows a $C\theta/T$ *versus* $(T/\theta)^2$ plot of the specific heat measurements on magnesium[28] and zinc.[22] Superimposed on these data are the straight lines predicted by the elastic constant measurements with the lines chosen to pass through the data at the low-temperature end. It is obvious that the line based on elastic data is not the best straight-line fit to the data and that a good description of the data occurs only below temperatures of about $\theta/150$ where we have required it to fit. In these cases, the elastic constants would have to be in error by about 4% in order to account for the difference in results. Such a difference is far outside experimental error for all the constants except C_{13}. Table III shows that if this one constant is changed by 1%, the value of θ is only changed by 0.1% which makes it unlikely that the inaccuracy of C_{13} could account for the discrepancy in θ values. It may be argued that the true Debye region has not been reached for these hexagonal metals even though the temperature is below $\theta/100$. However, the magnesium specific heat data does fall on a straight line which indicates that it can be well described with a T^3 dependence. The zinc data deviate slightly from a T^3 law and Seidel[22] has made a least squares fit to his heat capacity data assuming a temperature dependent θ. From this he extrapolates to $T=0°K$ to obtain a value of $\theta_0=309°K$ which is still 5% below the elastic constant value.

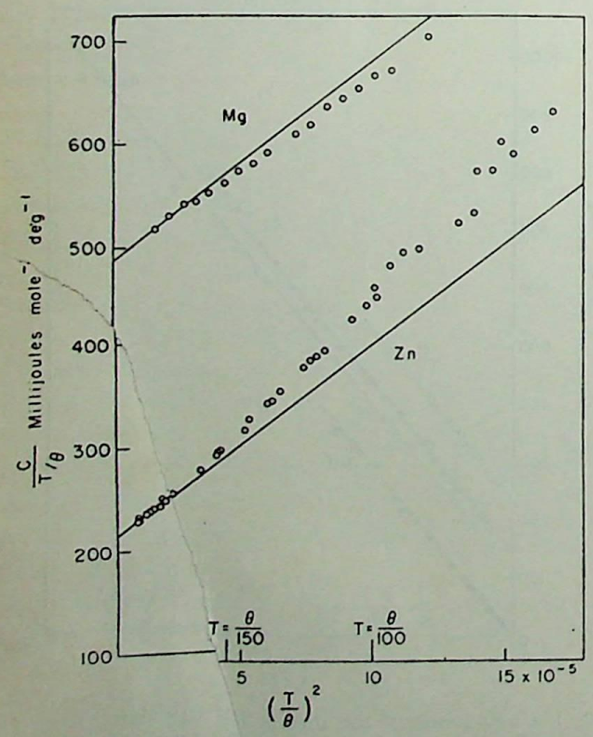


FIG. 2. Comparison of the specific heat calculated from Debye theory (solid line) with the measured specific heats of hexagonal metals.

TEMPERATURE DEPENDENCE OF θ

The previous section has shown the degree to which Debye theory describes experimental specific heat measurements near absolute zero. In order to calculate the lattice specific heat above $T=0$ using elastic constant data, it is necessary to modify Debye theory to include effects which arise because the solid is not a continuum but a lattice of discrete atoms. This requires considering a model in which the solid consists of mass points held in equilibrium by forces which are expressable in terms of the elastic constants. For convenience, the results of these modifications are usually expressed in terms of an effective θ which can be compared with the Debye θ. This modified θ is expected to vary with temperature and its temperature dependence can be calculated assuming a particular force constant model. From this, the lattice specific heat can be determined and compared to the calorimetric measurements in order to judge the validity of the model chosen. It is also possible to calculate a calorimetric θ from specific heat measurements and compare it with the modified θ predicted by the lattice theory. But unless very precise calorimetric measurements are used, the resulting values of θ may show considerable scatter and make a judgment of the theoretical model very difficult. We have chosen to compare the specific heat values themselves since these are of more technical interest and since it is then obvious how much error can be expected from using elastic constant models to estimate lattice specific heats.

Figure 3 shows the experimental data for copper,[32] silver,[33] and gold[34] plotted as C/T *versus* T^2. The solid curves labeled D show the variation of the specific heat predicted by the unmodified Debye theory which has the constant value of $\theta=\theta_0$. Gold, which shows deviations from the elastic constant θ_0 at temperatures of $\theta/60$, follows this theory to remarkably high temperatures ($T\sim\theta/7$) even though silver and copper deviate considerably from it at temperatures of the order of $\theta/20$. The curves marked L are based on the model of de Launay[1] who assumes a central force connecting nearest and next-nearest neighbor atoms

[32] W. F. Giauque and P. F. Meads, J. Am. Chem. Soc. 63, 1897 (1941).
[33] Meads, Forsythe, and Giauque, J. Am. Chem. Soc. 63, 1902 (1941).
[34] T. H. Geballe and W. F. Giauque, J. Am. Chem. Soc. 74, 2368 (1952).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

and a third force arising from the compressibility of the electron gas. This model does not describe the lattice specific heat for gold and copper over the temperature range in which data are available but seems to describe the silver data up to temperatures of approximately $\theta/10$.

The curves marked *H* are the predictions of the model of Horton and Schiff[35] which are based on assuming a three-force constant noncentral interaction between nearest neighbor atoms only. This model is in better agreement with the measurements for copper and gold, but not enough to be significant. None of the models gives a really satisfactory description of the specific heat in the range where measurements are available. However, this apparent failure may not be serious since these theories are expected to apply only up to temperatures of $\theta/20$ and the available data do not cover this range. A critical evaluation of the models must wait for accurate specific heat data between 4 and 15°K.

A model applicable to hexagonal metals has been calculated by Garland and Slutsky.[36] Using central forces between the first three nearest neighbors and a force due to the compressibility of the electron gas they find that θ decreases with increasing temperature but not as rapidly as is observed. They also find that zinc with its high c/a ratio cannot be described well with their model and do not attempt a comparison with the specific heat values.

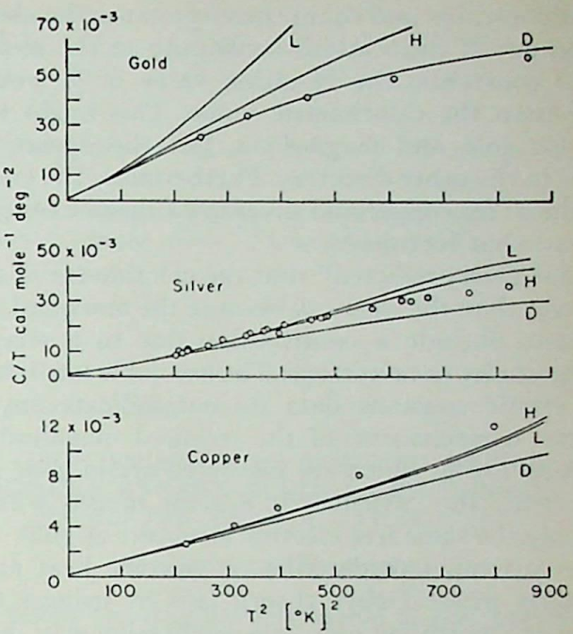


FIG. 3. Comparison of the specific heat calculated from several different models with the measurements in the 15 to 30°K range. The curves marked *D* are based on Debye's continuum model while those marked *L* and *H* refer to the lattice models proposed by de Launay and by Horton and Schiff, respectively.

CONCLUSIONS

For the cubic materials considered, the value of θ_0 calculated from elastic constants measured at 4°K by direct numerical integration is in very good agreement (except possibly for gold) with the low-temperature specific heat measurements. The tables of de Launay yield results which agree with the integration to within the accuracy of reading the table and is by far the easiest to apply. The two expansion methods of Betts *et al.* and of Quimby and Sutton appear to give θ_0 accurately to 1% even for the most anisotropic case but the method of Quimby and Sutton seems laborious and more susceptible to mistakes. Considering the differences in the results of the series approximation, and the errors in interpolation and extending de Launay's tables, we advance the values listed under the Numerical integration column in Table II to be the best available values calculated from elastic constants.

Since the lattice specific heat predicted by Debye theory from elastic constants is consistent with the calorimetric measurements to the extent shown in Fig. 1, it seems possible to conclude that the simple Debye model of lattice heat capacity is a very good approximation at low temperatures in cubic substances. In materials where the lattice specific heat cannot clearly be separated from other contributions (i.e., when there are large electronic specific heats, low-temperature phase transitions, or magnetic interactions) the elastic constants can be used to define the lattice contribution without introducing appreciable error.

For the hexagonal metals and gold there seems to be a definite difference between the calorimetric θ_0 and the elastic constant value. The reason for this is not certain. It is not likely that errors in elastic constants will account for the difference. Furthermore, for zinc two independently measured sets of elastic constants result in nearly the same θ_0 value. It is possible that small systematic errors in the specific heat measurements or uncertainties in the temperature scale would account for the difference.

It may be that temperatures of $\theta/60$ are not sufficiently low for the simple Debye model to apply to these metals. This may imply that dispersion of the elastic waves is an important effect at these temperatures. Simple reasoning indicates that for an isotropic medium deviations down from the straight line in a C/T *vs* T^2 plot are a result of an upward bending of the $\omega(k)$ curve and conversely for an upward deviation. Although not strictly applicable to anisotropic media, such reasoning would indicate that the $\omega(k)$ curve for gold and magnesium tends to bend upward away from the Debye approximation while for zinc it bends downward. However, examination of Figs. 1 and 2 indicates that the data need not be considered as "curving away" from the theoretical straight line since it can be fitted by a straight line of different slope. Thus the dispersion must be of a rather special nature.

There may be effects which modify the measured

[35] G. K. Horton and H. Schiff, Can. J. Phys. **36**, 1127 (1958).
[36] C. W. Garland and L. J. Slutsky, J. Chem. Phys. **28**, 331 (1958).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

elastic constants and could therefore cause the observed differences. If dislocations contribute to the measured elastic constants, the resulting value of θ_0 would be lower than the calorimetric value. This might be the case for gold and magnesium, but the deviation for zinc is in the other direction. Furthermore, the excellent agreement for copper and silver must then be considered as somewhat fortuitous.

It has been predicted[37] that the calorimetric θ_0 should be larger than the elastic θ_0 because the measured elastic constants include a contribution due to a relaxation mechanism of the electrons. The available low-temperature elastic constant data do not indicate any such relaxation mechanism of the required magnitude and the values of θ_0 which are predicted are in good agreement with the calorimetric θ_0's for metals with presumably the same free electron structure as gold.

The attempts to describe the specific heat at temperatures where Debye theory fails by using a lattice model are satisfying only in a qualitative way, but are unsatisfactory in predicting the detailed temperature dependence of θ. However, before a critical comparison can be attempted, accurate specific heat measurements must be made in the 4° to 20°K range.

ACKNOWLEDGMENTS

The authors are grateful to Dr. J. Zimmerman, Dr. A. Arrott, and Mr. J. Skalyo of this Laboratory, Dr. J. Rayne of the Westinghouse Research Laboratory, Mr. G. Seidel of Purdue University, Dr. C. W. Garland of MIT, and Dr. G. K. Horton and Dr. H. Schiff of the University of Alberta for communicating their results prior to publication. We also wish to thank Mr. M. T. MacNeil for his help in preparing the calculation program for the computer, and Dr. P. M. Sutton of the Corning Glass Works for calculating the values listed under his name in Table II.

ADDENDUM

Since submitting this article, more elastic constant and calorimetric data have become available. Calorimetric measurements by Phillips[38] down to $\theta/1000$ (0.3°K) gives $\theta_0=426°$ for aluminum which compares favorably with the calculated value of $\theta_0=428.2$ shown in Table II. New calorimetric measurement on LiF[39] down to $\theta/80$ give $\theta_0=722$°K which is in disagreement with the calculated value in Table II. However, these latter calorimetric data are somewhat scattered and, in fact, can be reasonably well described by a line corresponding to $\theta_0=734.1$°K.

Other new data are shown in the added tables. Additions to Table I shows the recent elastic constant data. The values of atomic volume and density are taken from the reference cited. Additions to Table II shows the resulting calculated values of θ_0.

As before, the calculated values of θ_0 are in good agreement with each other. The results of calorimetric measurements are also listed. For thorium, although the calorimetric θ_0 seems to be in disagreement, the calorimetric data up to $\theta/50$ is well described by choosing $\theta_0=164.2$°K. The calorimetric and elastic constant values for nickel are apparently in disagreement but the calorimetric data are quite scattered so that critical comparison is not possible.

For the two alkali halides listed, the calculated and calorimetric values are in apparent agreement. However, on a plot like Fig. 1 the elastic constant line through the origin does not fit the calorimetric data, which extends down to almost $\theta/100$. Dispersion effects can explain this failure as shown by Barron, Berg, and Morrison[40] who obtained the values of θ_0 listed above by extrapolating calorimetric data to $T=0$. This result indicates that deviations from the Debye approximation occur at relatively lower reduced temperatures for the alkali halides than for the cubic metals. If dispersion is to account for the discrepancies in the hexagonal metals, it must be concluded that the Debye approximation is inapplicable at reduced temperatures much lower than those which are suitable even for the alkali halides.

TABLE I. Additions.

	C_{11}	C_{12}	C_{44}	Ω^{-1}	ρ	Reference
Elastic constants extrapolated from 4.2°K						
KCl	0.483	0.054	0.0663	3.294	2.038	a
Nickel	2.612	1.507	1.317	9.206	8.968	b
Elastic constants extrapolated from 77°K						
KI	0.338	0.022	0.0368	2.321	3.197	a
Thorium	0.780	0.482	0.513	3.058	11.787	c

[a] M. H. Norwood and C. V. Briscoe, Phys. Rev. 112, 45 (1958).
[b] Alers, Neighbours, and Sato (to be published).
[c] Armstrong, Carlson, and Smith, J. Appl. Phys. 30, 36 (1959).

TABLE II. Additions.

	de Launay	Numerical integration	Calorimetric	Reference
Elastic constants extrapolated from 4.2°K				
KCl	236.3[e]	237.1	235.1	a
Nickel	476.3	476.2	441	b
Elastic constants extrapolated from 77°K				
KI	130.9[e]	131.5	132.3	a
			128	d
Thorium		164.2	170	e

[a] See reference 40.
[b] J. A. Rayne and W. G. Kemp, Phil. Mag. 1, 918 (1956).
[c] These values were calculated using tables prepared by J. de Launay which are to be published in J. Chem. Phys.
[d] See reference 39.
[e] P. L. Smith, Proc. Conf. de Phys. des Basses Temp., Paris, p. 283 (1955)

[37] J. L. Warren and R. A. Ferrell, Bull. Am. Phys. Soc. Ser. II, 3, 226 (1958).
[38] N. E. Phillips, *Proceedings of the Fifth International Conference on Low Temperature Physics and Chemistry* (University of Wisc. Press, Madison, Wisconsin, 1958), p. 414.
[39] W. W. Scales, Phys. Rev. 112, 45 (1958).
[40] Barron, Berg, and Morrison, Proc. Roy. Soc. (London) 242, 478 (1957).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

Masers*

J. WEBER

University of Maryland, College Park, Maryland

"It is strange that practical electronics remained untouched by these fundamental facts (electron spin) and could get along with the notion of the charged mass point or the minute charged sphere." [Arnold Sommerfeld, *Electrodynamics* (Academic Press, Inc., New York, 1948)].

I. INTRODUCTION

THE development of electronics during the first half of this century was based almost completely on the quantum states of translational energy of the free electron. Only recently did physicists recognize the fact that the spin states and the bound states of electrons in atoms and molecules could be employed for amplification. We give here an account of these technological advances together with a summary of the physics which made them possible.

The physical principles and experimental techniques required for development of paramagnetic amplifiers and molecular amplifiers were well established in the period 1945–1950. Such microwave amplifiers have much lower noise than those employing thermionic vacuum tubes. This low noise property has been the principal motivation for the large amount of work in this field.

The earliest work on molecular amplification started independently at three different laboratories.[1–3] It had a different motivation, namely, the hope that millimeter wave oscillators and amplifiers would result.

The word "maser" was coined by the Columbia group[2] as an acronym for "microwave amplification by the stimulated emission of radiation." This word appears in Webster's dictionary with the spellings "mazer" and "maser" and has the meaning "a large drinking cup, originally of a hard wood." The free electron vacuum tube amplifier may also be regarded as operating thfough the mechanism of the stimulated emission of radiation. In what follows, however, the word maser is taken to mean either a molecular amplifier or a paramagnetic solid state amplifier.

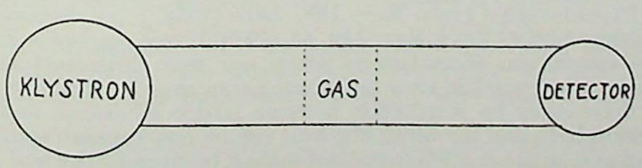


FIG. 1. Absorption of microwaves by a gas.

II. AMPLIFICATION BY SYSTEMS HAVING A HIGHER ENERGY STATE MORE DENSELY POPULATED THAN A LOWER STATE

A. General Principles

We follow the original[1] discussion of this principle of maser type amplification. In the experimental arrangement of Fig. 1, there is a microwave system which consists of a wave guide with gas inside, in thermal equilibrium. Such a gas may absorb microwaves, and this absorption can be described in the following terms.

Suppose the gas has a pair of energy levels (Fig. 2) with values E_1 and E_2, with $E_2 > E_1$. Let n_1 and n_2 be the numbers of particles with energies E_1 and E_2. Let the absolute temperature be denoted by T and Boltzmann's constant by k. Then we can write

$$n_2 = n_1 e^{-(E_2-E_1)/kT}. \quad (1)$$

We suppose that electromagnetic radiation is present with frequency ν given by

$$\nu = (E_2-E_1)/h. \quad (2)$$

The power absorbed by the gas can be written

$$P_A = W_{12} n_1 h\nu. \quad (3)$$

Here W_{12} is the transition probability for transitions from state 1 to 2 induced by radiation. Similarly the power emitted by the gas due to the stimulated emission of radiation is given by

$$P_E = W_{21} n_2 h\nu. \quad (4)$$

Neglecting spontaneous emission for the moment we can write $W_{21} = W_{12}$ and for the *net* power absorbed

$$P = W_{12}(n_1-n_2)h\nu. \quad (5)$$

P will be positive and the gas will absorb power, if n_1 exceeds n_2. This will be the case in consequence of expression (1), if the gas is in thermal equilibrium. How-

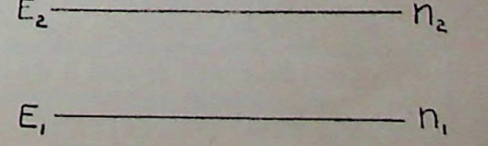


FIG. 2.

* This work was supported in part by the Office of Naval Research.

[1] J. Weber, Trans. Inst. Radio Engrs. Prof. Group on Electron Devices, PGED-3 (June, 1953).

[2] Gordon, Zeiger, and Townes, Phys. Rev. **95**, 282 (1954); Phys. Rev. **99**, 1264 (1955).

[3] N. G. Basov and A. M. Prokhorov, J. Exptl. Theoret. Phys. (U.S.S.R.) **27**, 431 (1954); Proc. Acad. Sci. (U.S.S.R.) **101**, 47 (1955); J. Exptl. Theoret. Phys. (U.S.S.R.) **28**, 249 (1955).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

ever, the net power absorbed can be made negative, and amplification will result, if $n_2 > n_1$. For the present we consider amplification by a one-quantum process, resulting from the nonequilibrium situation $n_2 > n_1$. The same considerations which lead to coherent absorption for $n_1 > n_2$ suggest that for[4] $n_2 > n_1$ the amplified signal will be coherent with the driving signal. The energy levels which are employed may be bound states of molecules, or spin states in a magnetic field. The low-temperature solid state masers use paramagnetic ions in an externally applied magnetic field.

B. Three-Level Method

Many methods have been proposed for obtaining a higher population density in an upper energy level than in a lower one. Most of these are not generally useful, but may nonetheless have special applications. A fairly complete survey is given in Sec. IV. First we present a summary of the three-level method, first proposed by Prokhorov,[3] and developed independently by Bloembergen,[5] and also by Javan; this appears to be by far the most generally useful one. We follow Bloembergen's treatment.

In a system with three energy levels, E_1, E_2, and E_3 (Fig. 3), let the selection rules be such that transitions are allowed between each level and either of the other two. In thermal equilibrium, the numbers of particles in the different states satisfy the relations,

$$n_1 > n_2 > n_3. \tag{6}$$

The frequencies ν_{32}, ν_{21}, ν_{31} are defined by the relations

$$\nu_{mn} = (E_m - E_n)/h. \tag{7}$$

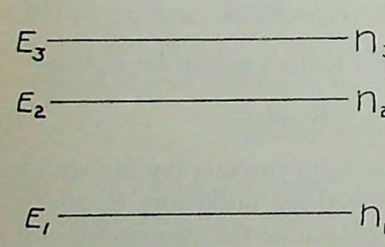


FIG. 3.

An intense rf field of frequency ν_{31} will induce transitions between states 1 and 3. Since initially there are more particles in state 1, the particles will initially leave state 1 at a greater rate than return to it from state 3.

Saturation phenomena will result in which $n_3 \to n_1$. Under these nonequilibrium conditions we may expect either that $n_2 > n_1$ or $n_3 > n_2$. In the first case amplification is possible at frequency ν_{21} and in the second case at ν_{32}. In order to arrive at quantitative values relaxation[6]

[4] The term negative temperature [E. M. Purcell and R. V. Pound, Phys. Rev. 81, 279 (1951)] is often used to describe a system in which higher energy levels are more densely populated than lower ones. This is consistent with expression (1). The idea of a negative temperature implies that while part of a system is not in equilibrium with its environment, it is sufficiently isolated from its positive temperature bath that it can be considered as a separate thermodynamic system described by a temperature. The concept of spin temperature has been carefully considered by A. Abragam and W. G. Proctor [Phys. Rev. 109, 1441 (1958)]. Negative temperature is only possible for a system whose energy levels have an upper bound. Thus when such a system is heated from 0°K to a temperature $+\infty$ the levels become equally populated. If it is heated further the temperature changes discontinuously to $-\infty$, and on further heating it eventually tends to minus 0°K at which point all particles are in the state of highest energy. As Bloembergen has remarked, negative temperature is hotter than hot. It has been suggested by Ramsey that the discontinuity at $+\infty$ could have been avoided if the reciprocal of temperature had been used. At negative temperatures resistances become negative and amplification is then possible.

[5] N. Bloembergen, Phys. Rev. 104, 324 (1956).

[6] An important issue in the operation of a paramagnetic ion low-temperature maser is the mechanism of paramagnetic relaxation. We discuss briefly some pertinent aspects of this problem. It appears more complex than the corresponding nuclear magnetic resonance problem. One experimental fact is that it is possible to couple the saturation radiation to one set of spin levels and simultaneously saturate a second pair of levels at approximately the same spacing which is weakly coupled to the radiation field [Strandberg, Davis, Faughnan, Kyhl, and Wolga, Phys. Rev. 109, 1988 (1958)]. Strandberg [Phys. Rev. 110, 65 (1958)] has interpreted these results as evidence for the idea that there are anomalies in the phonon excitation, with the lattice modes with frequency near the spin resonance in equilibrium with the high-temperature saturated spin system. He considers this to be in part a consequence of the fact that the specific heat of the lattice vibrations at 4°K (calculated from say the Debye model) is several orders smaller than that of the paramagnetic spin system. Giordmaine, Alsop, Nash, and Townes [Phys. Rev. 109, 302 (1958)] have described a series of experiments using paramagnetic resonance at 9000 Mc, at 1–4°K; $Gd_2Mg_3(NO_3)_{12}\cdot 24H_2O$), $K_3Cr(CN)_6$, and $Cu(NH_4)_2(SO_4)_2\cdot 6H_2O$) were used. They found that it was not possible to "burn a hole" in their resonances by saturation, indicating that their lines were homogeneously broadened. In the case of Cu, saturation of one line immediately saturated seven neighboring resonances. They were not able to observe the spin reversal by adiabatic rapid passage. No difference in absorption was noted in times as short as one msec after adiabatic rapid passage. They also observed that excitation due to a rapid sweep through a single resonance decayed very rapidly while that due to a slower sweep decayed at the normal rate, $T_1 \to 10$ sec. They interpreted these results as evidence for the idea that for some salts at low temperatures it is the lattice bath relaxation which limits the total relaxation rate. [Gorter, Van der Marel, and Bölger, Physica 21, 103 (1955)]. They also remarked that for these salts the spin lattice relaxation time is several orders smaller than the observed values of T_1, and is of order 10^{-5} sec for Gd and 10^{-4} sec for Cu. In addition, they concluded that the breadth of the lattice modes is much larger than the width of the resonances in the diluted crystals (several hundred Mc for 1% paramagnetic concentration of the Cu salt), that the relaxation time T_1 is dependent on crystal size, and that the breadth of the lattice modes increases with increasing concentration of paramagnetic centers. They also suggested that for those salts described by the foregoing conditions, operation of an adiabatic rapid passage two-level maser would be impractical while three-level maser operation would be possible with reduced band width.

Bloembergen [Phys. Rev. 109, 2209 (1958)] and Strandberg independently [Phys. Rev. 110, 65 (1958)] have reached a different conclusion. Bloembergen points out that three-level maser iction is not possible for a salt whose paramagnetic relaxation rate is determined by interaction between lattice vibrations and the helium bath and for which the heat conduction between spin system and lattice is 1000 times better than between lattice and heat bath. Let ν_{31} be the saturation frequency and ν_{32} the amplification frequency and let these frequencies be widely spaced, with no overlapping of phonon bands. Then we have a negative spin temperature corresponding to levels 2 and 3. If the thermal contact with the phonons is very good then a band centered at frequency ν_{32} will start to gain energy, in order to attempt to reach equilibrium with the negative temperature spin system. Since the lattice vibration levels are those of harmonic oscillators, they have no upper bound. They cannot be saturated [R. Karplus and J. Schwinger Phys. Rev. 73, 1020 (1948)] and the temperature associated with the ν_{32} band can never become negative. The steady state (nonequilibrium) situation would be one in which the spins of levels 2 and 3 are at a very high negative temperature and those

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

effects must be included. Let w_{12} be the heat bath induced transition probability from state 1 to state 2, with corresponding meanings for w_{21}, w_{32}, w_{23}, w_{13}, and w_{31}. First suppose the system is in thermal equilibrium with no microwave fields applied. The number of particles leaving the state E_1 per sec must equal the number returning to it, so

$$n_1w_{12}=n_2w_{21}, \quad n_1w_{13}=n_3w_{31}. \tag{8}$$

Employing the Boltzmann factor enables us to write

$$w_{12}/w_{21}=n_2/n_1=e^{-h\nu_{21}/kT}. \tag{9}$$

The w's are the reciprocals of the spin lattice relaxation times. Now suppose that a strong microwave field of frequency ν_{31}, and a weak microwave signal of frequency ν_{32} are present. Let W_{31} and W_{32} be the transition probabilities induced by these rf fields. The numbers of particles n_1, n_2, and n_3 occupying the three levels satisfy the relation,

$$n_1+n_2+n_3=N. \tag{10}$$

For $(h\nu_{32})/(kT)\ll 1$, the populations satisfy the equations

$$\frac{dn_3}{dt}=w_{13}\left(n_1-n_3-\frac{Nh\nu_{31}}{3kT}\right)+w_{23}\left(n_2-n_3-\frac{Nh\nu_{32}}{3kT}\right)$$
$$+W_{31}(n_1-n_3)+W_{32}(n_2-n_3),$$

$$\frac{dn_2}{dt}=w_{23}\left(n_3-n_2+\frac{Nh\nu_{32}}{3kT}\right)$$
$$+w_{21}\left(n_1-n_2-\frac{Nh\nu_{21}}{3kT}\right)+W_{32}(n_3-n_2), \tag{11}$$

$$\frac{dn_1}{dt}=w_{13}\left(n_3-n_1+\frac{Nh\nu_{31}}{3kT}\right)$$
$$+w_{21}\left(n_2-n_1+\frac{Nh\nu_{21}}{3kT}\right)-W_{31}(n_1-n_3).$$

These equations have the following approximate steady-state solution for the case in mind, $W_{31}\gg W_{32}$.

$$n_1-n_2=n_3-n_2=\frac{hN}{3kT}\left[\frac{-w_{23}\nu_{32}+w_{21}\nu_{21}}{w_{23}+w_{21}+W_{32}}\right]. \tag{12}$$

In case the numerator of (12) is positive, amplification is possible at frequency ν_{32}. In case it is negative, amplification is possible at the frequency ν_{21} and the W_{32} in the denominator of (12) must be replaced by W_{21}. The signal power emitted by the material is obtained by multiplying (12) by the signal induced transition probability W_{32} and the energy of one quantum.

$$P=\frac{Nh^2\nu_{32}(w_{21}\nu_{21}-w_{32}\nu_{32})W_{32}}{3kT(w_{23}+w_{21}+W_{32})}. \tag{13}$$

The situation described by (12) and (13) is qualitatively the same, but more complex, if $(h\nu/kT)$ is not small compared to 1. The procedures for calculating the transition probability W_{32} require a knowledge of the matrix elements of the interaction with the Maxwell field, and appropriate relaxation times. For the remainder of the section we consider a paramagnetic ion type of solid state amplifier. In this case the signal induced transition probability is given by

$$W_{32}=\left(\frac{2\pi}{h}\right)^2|\langle 2|M_x|3\rangle|^2\langle H_s^2(\nu_{32})\rangle_{Av}T_2. \tag{14}$$

Here M_x is the (x component) magnetic dipole operator, $\langle H_s^2(\nu_{32})\rangle_{Av}$ is the volume average squared magnetic field at the signal microwave frequency ν_{32}, and T_2 is the spin-spin relaxation time.

Expression (13) shows that the temperature T of the heat bath must be low. The three-level solid state masers which have so far been successfully operated have employed liquid helium cooling with temperatures in the range 1.25–4°K. In addition to increasing the amplification, the low temperatures partially improve the noise performance.

The power which is absorbed from the saturation†

lattice vibrations with which the spins are on "speaking terms" will be at a very high positive temperature. A high negative temperature means that the level populations associated with states 2 and 3 are almost equal, and therefore the maser would not operate in the observed manner. Bloembergen has solved the problem of three-level maser and phonon steady states in terms of the spin populations n_1, n_2, n_3 and the average lattice oscillator excitation quantum numbers in the three phonon bands, $n_{ph}(\nu_{32})$, $n_{ph}(\nu_{21})$, $n_{ph}(\nu_{31})$. His solution confirms that n_3-n_2 is too small for maser operation if the contact between spins and phonons is better than between phonons and heat bath.

A possibility which has briefly been explored by the author is that the temperature associated with all three maser spin levels is positive, amplification occurring as a result of a two-quantum process involving absorption of a saturation frequency photon and emission of an amplification frequency photon. This is a kind of Raman process in which energy is conserved in both transitions. Direct calculation shows that this effect is too small, and that a low negative spin temperature needs to be assumed, in conjunction with single photon processes, in order to explain the operation of a maser.

It can be concluded that the paramagnetic relaxation mechanisms suggested by Giordmaine, Nash, and Townes do not apply to those salts which have been successfully used in a three-level maser.

Shapiro, Bloembergen, and Artman [Bull. Am. Phys. Soc. Ser. II, 3, 317 (1958)] have reported additional indirect saturation experiments which they interpret as caused by higher order spin-spin interactions rather than by "hot phonons." This appears very reasonable because it is difficult to see how the lattice vibration modes could be tightly coupled to the saturated spin system without being similarly coupled to the negative temperature spin system. Further work by Bloembergen, Shapiro, Pershan, and Artman (Cruft Laboratory Technical Report No. 285, Harvard University, Cambridge, Massachusetts, October 15, 1958) has increased the evidence that the indirect saturation phenomena are indeed due to spin spin interactions. This follows the mechanism proposed by R. Kronig and C. J. Bouwkamp, Physica 5, 521 (1938) and Physica 6, 290 (1939).

† *Note added in proof.*—An interesting discussion of masers as heat engines has been given by H. E. D. Scovil and E. O. Schulz-

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

field of frequency ν_{31} is given by that needed to balance the tendency of collisions to restore equilibrium, and is

$$P_s=(h\nu_{31})^2 n_1/2T_1kT. \quad (15)$$

Here T_1 is the spin-lattice relaxation time.

Expressions (13), (14), and (15) lead to the following general criteria for selection of materials for a paramagnetic ion three-level maser. If the heat bath transition probabilities w_{21} and w_{23} are nearly equal, the frequencies ν_{21} and ν_{32} should be very unequal. However, if the frequencies ν_{21} and ν_{32} are approximately equal, then w_{21} and w_{23} should be very unequal. If ν_{21} is exactly equal to ν_{32} the device cannot operate since the same signal which induces transitions between states 2 and 1 will be effective in inducing transitions between 3 and 2. Amplification between one pair of states would be annulled by absorption associated with the other pair since (12) requires that $n_1-n_2\approx n_3-n_2$. In order to obtain large gain the temperature T has to be small. The spin should be at least 1, but preferably not more than $\frac{3}{2}$, otherwise the particles will be distributed among too many states. (Gd with a spin of $\frac{7}{2}$ is a notable exception.) It is also desirable that the nuclear spin be zero. Expression (15) requires that for small T, the spin-lattice relaxation time T_1 be as large as possible, otherwise excessive saturation power will be needed with consequent difficulty in maintaining low temperatures. Expressions (13) and (14) show that the product of total number of spins, N, and the spin-spin relaxation time T_2 should be as large as possible. Inasmuch as an increase in concentration of spins tends, in general, to decrease T_2 it is necessary to attain an optimum value of the product NT_2 by suitably diluting the paramagnetic ions. The zero field splitting ought to be of the same order of magnitude as the energy level difference associated with the signal frequency. This follows because mixing of the spin states is essential in order to have transitions allowed between all three levels. This is most favorable when the Zeeman and crystalline field terms are comparable.

It is possible, and in some cases desirable, to use a four-level system. A particularly elegant solution of the problem of four-level maser design employing ruby has been given by Kikuchi,[7] Makhov, Lambe, and Terhune with $E_4-E_2=E_3-E_1$. In the latter case a pumping frequency $\nu_{42}=\nu_{31}$ populates the third level and simultaneously depopulates the second level.

A generalization of (13), given by Kikuchi, is

$$P=\frac{Nh^2}{4kT}\left[\frac{w_{21}\nu_{21}-w_{32}\nu_{32}+w_{41}\nu_{41}+w_{43}\nu_{43}}{w_{21}+w_{23}+w_{14}+w_{34}+W_{32}}\right]W_{32}\nu_{32}. \quad (16)$$

The active material of a maser can be placed in a wave-guide transmission system or in a resonant cavity. In the former case we have a traveling wave amplifier with power output given by

$$P=P_oe^{\beta l}, \quad (17)$$

where l is the distance along the amplifier.

Differentiating this with respect to l leads to the gain coefficient β given by

$$\beta=dP/Pdl. \quad (18)$$

Here (dP/dl) is the power emitted per unit length and P is the power at the point where (dP/dl) is calculated, in accordance with expressions (13) or (16). Employing (14), this can be written

$$\beta=\frac{32\pi^3(n_3-n_2)\,|M_{32}|^2T_2\nu f}{hv_g}, \quad (19)$$

where f is a filling factor which may approach 1 and v_g is the group velocity.

Here the power P has been set equal to the energy density times the group velocity v_g. This equation shows that a slow wave structure (small group velocity) is a desirable method of obtaining a higher gain-band-width product. Slow wave structures proposed thus far are a ruby rod with a helix wrapped around it, considered by W. W. Anderson of Stanford University, and a rectangular wave guide with conducting fingers giving circular polarization, considered by DeGrasse, Schulz-DuBois,

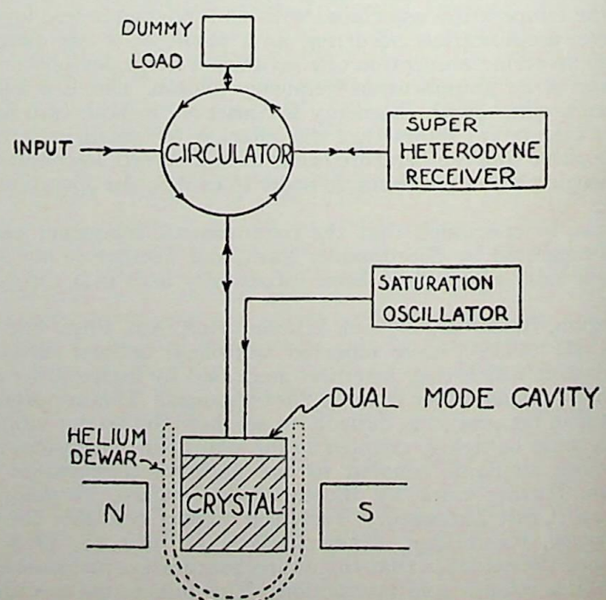


FIG. 4. Reflection cavity maser.

DuBois [Phys. Rev. Letters 2, 262 (1959)]. They consider a hot reservoir with temperature T_1 and filter allowing a band of frequencies in the vicinity of ν_{31} to pass, in thermal contact with a maser. Levels 2 and 3 are in thermal contact with a cold reservoir at temperature T_0, and coupled through a filter with pass band centered about ν_{32}. The signal frequency is ν_{21}. Ignoring relaxation processes within the maser, using the Boltzmann distribution law and requiring $n_2>n_1$ leads to the result that the maser efficiency ν_{21}/ν_{31} is that of the Carnot engine, $(T_1-T_0)/T_1$. They note the important result that instead of using a coherent microwave pump at ν_{31}, thermal excitation by two reservoirs as described above should make it possible to generate microwaves.

[7] Makhov, Kikuchi, Lambe, and Terhune, Phys. Rev. **109**, 1349 (L) (1958); also Kikuchi, Lambe, Makhov, and Terhune (to be published).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

and Seidel of the Bell Telephone Laboratories. These structures could be employed either as transmission or reflection devices. Means must be provided to prevent output power from being fed back to the input, causing the amplifier to oscillate, and for preventing noise from the load from getting to the maser and being amplified. Transmission devices utilizing nonreciprocity would accomplish this.

The reflection cavity type of maser has undergone most development and appears capable of giving enough gain-band-width product for most purposes. A block diagram is sketched in Fig. 4. Here a nonreciprocal device, the microwave circulator,[8] prevents power from being fed back and causing oscillation, and keeps noise from the load out of the maser. Resonant cavity expressions for gain and noise performance are usually expressed in terms of the Q's (quality factors) of a cavity. The Q is defined by

$$Q=2\pi\nu E/P_a, \tag{20}$$

where E is the energy in the principal cavity mode and P_a is the rate of absorption of energy. The unloaded Q, usually denoted by Q_o, is defined by

$$Q_o=2\pi\nu E/P_{aw}, \tag{21}$$

where P_{aw} is the part of the absorbed power that is absorbed in the cavity walls. The magnetic Q, denoted by Q_m, is defined by

$$Q_m=-\frac{2\pi\nu E}{P_e}=\frac{-\nu\langle H^2(\nu)\rangle V_c}{4P_e}, \tag{22}$$

where P_e is the net power emitted by the active material.

Here V_c is the volume of the cavity and ν is the frequency being amplified. The denominator is given by expressions (13) or (16). The external Q, denoted by Q_e, is defined by

$$Q_e=2\pi\nu E/P_{en}, \tag{23}$$

where P_{en} is the net power output.

The cavity mode may be represented by the equivalent circuit in Fig. 5. G_m, G_o, and G_e are conductances associated with the spin system, the walls of the cavity, and the output load, respectively. From Fig. 5, a loaded Q may be defined as

$$Q_l^{-1}=Q_m^{-1}+Q_o^{-1}+Q_e^{-1}. \tag{24}$$

The voltage standing wave ratio in the transmission system which drives the cavity, B, is given in terms of

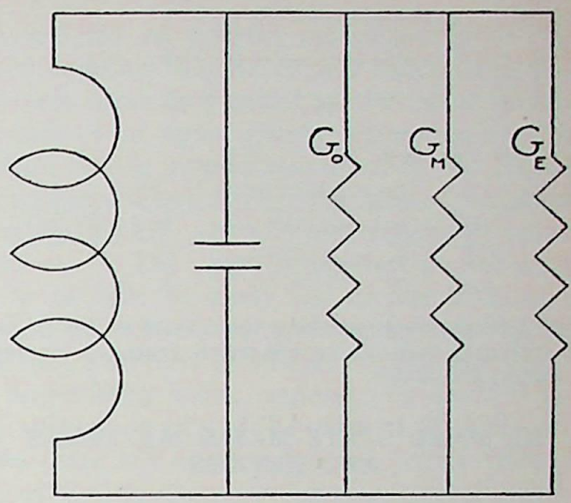


FIG. 5. Cavity equivalent circuit.

the incident voltage V_i and the reflected voltage V_r, as

$$B=\frac{|V_i|+|V_r|}{|V_i|-|V_r|}=\frac{1+g}{1-g}. \tag{25}$$

Here g is the "voltage" gain defined by $g=|V_r/V_i|$. The power gain is g^2.

After solving (25) for g^2, we obtain

$$g^2=[(1-B)/(1+B)]^2. \tag{26}$$

The maser will oscillate if $-Q_m^{-1}>Q_o^{-1}+Q_e^{-1}$, and will amplify if $Q_o^{-1}+Q_e^{-1}>-Q_m^{-1}>Q_o^{-1}$. Let us consider some characteristics of an amplifier.

From transmission line theory, the voltage standing wave ratio B (at resonance), in terms of the conductances shown in Fig. 5, is

$$B=(G_m+G_o)/G_e=Q_e[Q_o^{-1}+Q_m^{-1}]. \tag{27}$$

This assumes that the input transmission line is matched to a conductance (G_m+G_o).

Use of (26) gives for the power gain,

$$g^2=\left[\frac{Q_e^{-1}-(Q_o^{-1}+Q_m^{-1})}{Q_e^{-1}+(Q_o^{-1}+Q_m^{-1})}\right]^2. \tag{28}$$

In terms of the loaded Q this takes the form

$$g^2=[2Q_lQ_e^{-1}-1]^2. \tag{29}$$

The band width is obtained by dividing the operating frequency by the loaded Q. For a low-temperature maser Q_o is so large that it may be neglected in (28). The product of the square root of power gain and band width is

$$g\Delta\nu=\frac{\nu[Q_e+|Q_m|]}{Q_e|Q_m|}. \tag{30}$$

For a high-gain device it is customary to adjust Q_e so that it equals the magnitude Q_m. Under these conditions (30) is approximately a constant, given by

$$g\Delta\nu\approx 2\nu/|Q_m|. \tag{31}$$

[8] Microwave circulators are not available below about 1400 mc/sec. Autler has proposed an ingenious arrangement [Lincoln Laboratory, MIT, Rept. M 37-27; Proc. Inst. Radio. Engrs. 46, 1880 (1958)] using two masers in a "balanced" arrangement, which does not require use of a circulator. The two masers are at opposite ends of a coaxial or wave-guide "magic" tee. One of the maser arms is a quarter wavelength longer than the other. The antenna and receiver are connected to the other two arms. Thus the two maser outputs combine at the load. Noise from the load is amplified by the masers but is then radiated back out through the antenna.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

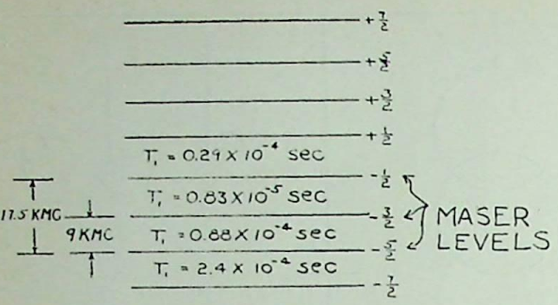


FIG. 6. Energy levels of the ground state of Gd^{+++} in ethyl sulfate, including spin-lattice relaxation times, as employed by Scovil, Feher, and Seidel.

III. SOLID STATE MASER MATERIALS AND DEVICES

The materials which have so far been successfully employed in three-level masers are[9,10] gadolinium ethyl sulfate, potassium chromicyanide,[11,12] and ruby.[7] We use the roman type g for the magnitude of the Zeeman tensor. This should not be confused with the italic g used for voltage gain.

Gadolinium ethyl sulfate was used in the first successful solid state maser.[10] The gadolinium ion is in an ^8S ground state having 7 electrons in a half-filled $4f$ shell. There is fine structure splitting into 7 lines with spacing which varies approximately as $3\cos^2\theta - 1$, where θ is the angle between the constant magnetic field, H_o, and the crystalline field axis. If the steady magnetic field is normal to the crystalline symmetry axis, the spin Hamiltonian may be written[11] approximately as

$$\mathcal{H} = g\beta_m H_o \cdot S_z - \tfrac{1}{2}D[S_z^2 - \tfrac{1}{3}S(S+1)] + \tfrac{1}{2}D[S_x^2 - S_y^2], \quad (32)$$

where β_m is the Bohr magneton, S is the spin operator, $g = 1.99$, $D = .02$ cm^{-1}, and the axis of quantization is parallel to that of the constant magnetic field, H_o. The first term of (32) represents the interaction with the field H_o which brings the transition to the required frequency. The second term makes the level spacings unequal, and the third term mixes the states, giving rise to transitions with $\Delta S_z = \pm 2$. The angle between H_o and the microwave magnetic field should be zero for the $\Delta S_z = \pm 2$ transitions and 90° for the $\Delta S_z = \pm 1$ transitions. Scovil, Feher, and Seidel used an angle of 45°. The energy levels and relaxation times of the ground state of Gd^{+++} in ethyl sulfate are shown in Fig. 6. Because the energy level separations are almost equal,

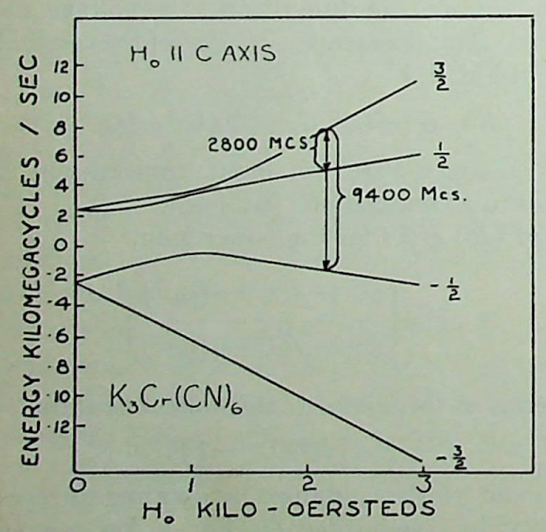


FIG. 7. Energy levels of $K_3Cr(CN)_6$ in $K_3Co(CN)_6$ *vs* applied magnetic field, for $H_o \parallel$ to the c axis.

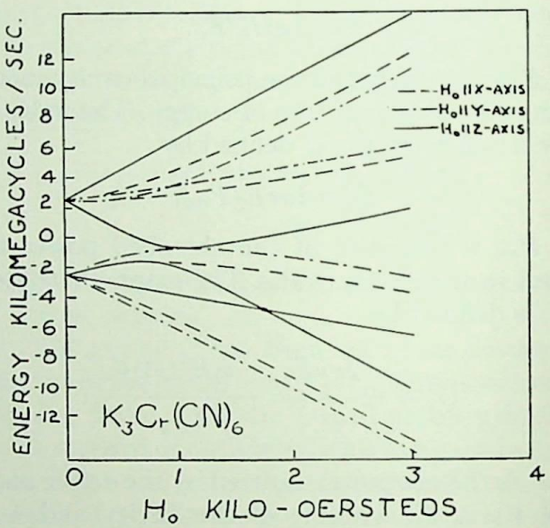


FIG. 8. Energy levels of $K_3Cr(CN)_6$ in $K_3Co(CN)_6$ *vs* applied magnetic field for H_o parallel to x, y, and z axes.

expression (13) requires that the spin-lattice relaxation times w_{21}^{-1} and w_{32}^{-1} be very unequal. A value for w_{32} ten times that of w_{21} is accomplished by introduction of cerium[9] into the crystal. An optimum value of NT_2 was obtained by employing a 90-mg (8% of the resonant cavity effective volume was filled) lanthanum ethyl sulfate crystal containing 0.5% Gd^{+++} and 0.2% Ce^{+++}. The device was operated at 1.2°K with a saturation frequency of 17.5 kmc/sec. Oscillations were obtained at 9 kmc/sec.

The second material to be successfully used[12] in a solid state maser was potassium chromicyanide diluted in a cobalticyanide crystal [$K_3Co(CN)_6$]. In this case 0.5% Cr was used as the paramagnetic salt. This material appears to have the long spin-lattice relaxation time of 0.2 sec, at 1.25°K. The paramagnetic resonance spectrum of $K_3Cr(CN)_6$ arises from two differently oriented complexes per unit cell. The spin Hamiltonian is

$$\mathcal{H} = \beta_m \mathbf{H} \cdot \mathbf{g} \cdot \mathbf{S} + D[S_z^2 - \tfrac{1}{3}S(S+1)] + E[S_x^2 - S_y^2]. \quad (33)$$

[9] G. Feher and H. E. D. Scovil, Phys. Rev. **105**, 2, 760(L) (1957).
[10] Scovil, Feher, and Seidel, Phys. Rev. **105**, 2, 762(L) (1957); H. E. D. Scovil, Trans. Inst. Radio Engrs. Prof. Group on Microwave Theory and Techniques **6**, 29 (1958).
[11] B. Bleaney and K. W. H. Stevens, Repts. Progr. in Phys. **16**, 108 (1951).
[12] A. L. McWhorter and J. W. Meyer, Phys. Rev. **109**, 312 (1958); R. H. Kingston, Lincoln Laboratory, MIT, Rept. M 35-79; Artman, Bloembergen, and Shapiro, Phys. Rev. **169**, 1392 (1958); S. H. Autler and N. McAvoy, Phys. Rev. **110**, 280(L) (1958).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

For cobalt as the diluent $D=0.083$ cm^{-1}, $E=0.011$ cm^{-1}, and the Zeeman tensor $\mathbf{g}$ is approximately isotropic and equal to 1.99. The direction cosines between the magnetic axes (x, y, and z) and the pseudo-orthorhombic crystalline axes (a, b, and c) are

	x	y	z
a	0.104	0	0.994
b	$\pm$0.994	0	$\mp$0.104
c	0	1	0

Energy level diagrams computed by McWhorter and Meyer are shown in Figs. 7 and 8. They used a dual mode coaxial cavity one half wavelength long at 2800 mc/sec operating in the TEM mode. For pumping, the cavity operated in the TE_{113} mode. The constant magnetic field H_o was applied, approximately parallel to the crystalline c axis, with operating conditions shown in Fig. 7. The upper two levels, labeled $+\frac{3}{2}$ and $+\frac{1}{2}$ were used for amplification and the second and fourth levels

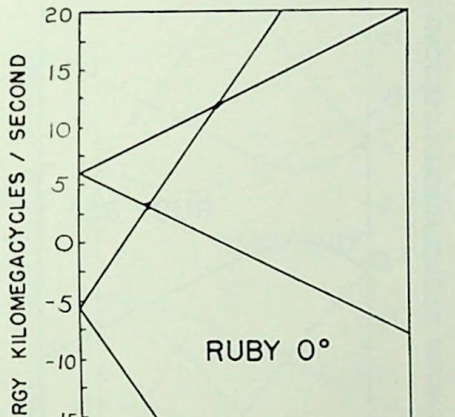


FIG. 9. Energy levels of Cr^{+++} in ruby *vs* applied magnetic field for H_o parallel to c axis of the crystal.

(labeled $+\frac{3}{2}$ and $-\frac{1}{2}$) were saturated by the rf field. Figure 7 shows that the energy level separation of the first and second levels (labeled $-\frac{3}{2}$ and $-\frac{1}{2}$) is approximately the same as that of the second and fourth levels. It has been shown[13] that for the Meyer-McWhorter arrangement (constant magnetic field parallel to the c axis and rf magnetic field parallel to the a axis) the transition probability for the first- to second-level transition is in fact 100 times greater than for the desired second and fourth levels. This is because the matrix elements are 4 times greater and the saturation radio-frequency magnetic field normal to the constant magnetic field favors the $\Delta m=1$ transitions. This implies that indirect saturation was accomplished by McWhorter and Meyer. They coupled strongly to levels 1 and 2 and succeeded in saturating levels 2 and 4 also. This possibility of indirect saturation of one pair of levels by coupling to another pair which have the same energy separation is a useful aid in maser design. The total number of spins in the upper quartet of the chromium ion is approximately 10^{19}. A 10% filling factor was used. The calculated gain-band-width product $g\Delta\nu$, was 2.6×10^6 sec^{-1}. The measured value was 1.8×10^6 sec^{-1}. The small filling factor was chosen in order to be able to study the device without undue distortion of the cavity electromagnetic mode configurations. Practical amplifiers would therefore use a much larger filling factor, approaching unity. This gives a large increase in $g\Delta\nu$, of the order of 50. This is partly because there are more spins and partly because the increased dielectric constant enables a smaller cavity to be used.

The third material to be successfully used was ruby. Paramagnetic resonance in this material was investigated in the Soviet Union in 1955 and 1956,[14] and in the United States by Geusic.[15]

Again the Cr^{+++} ion is used, but the host crystal is Al_2O_3. The four nonequivalent sites are indistinguishable since the spin is less than two, and the crystal has trigonal symmetry. Maser action in ruby was first demonstrated by Makhov, Kikuchi, Lambe, and Terhune at the University of Michigan.[7] The spin Hamiltonian for Cr^{+++} in the Al_2O_3 lattice is given by

$$\mathcal{H}=\beta_m\mathbf{S}\cdot\mathbf{g}\cdot\mathbf{H}+D[S_z^2-\tfrac{1}{3}S(S+1)]. \qquad (34)$$

$\mathbf{g}$ is the Zeeman tensor with components $g_{zz}=g_{\parallel}$, $g_{xx}=g_{yy}=g_{\perp}$. This Hamiltonian leads to the eigenvalue equation[7]

$$(\epsilon^2-1)^2-\tfrac{5}{2}x^2\epsilon^2+\tfrac{9}{16}x^4-\frac{x^2(5g_{\parallel}^2\cos^2\theta-g_{\perp}^2\sin^2\theta)}{2g^2}$$
$$+\frac{2\epsilon\lambda x^2(g_{\perp}^2\sin^2\theta-2g_{\parallel}^2\cos^2\theta)}{g^2}=0. \qquad (35)$$

Here

$$\epsilon=\frac{\text{Energy}}{|D|},\quad x=g\frac{\beta_m H}{|D|},$$

$$D=-0.1913\text{ cm}^{-1}=3.798\times10^{-17}\text{ ergs},$$

$$\lambda=\frac{D}{|D|},\quad g=1.986,\quad g^2=g_{\parallel}^2\cos^2\theta+g_{\perp}^2\sin^2\theta.$$

This equation is readily solved for the situations $\theta=0$, $\theta=\pi/2$, $g_{\perp}^2\sin^2\theta=2g_{\parallel}^2\cos^2\theta$. In the latter case, $\theta=\cos^{-1}(1/\sqrt{3})$ gives (assuming isotropic $\mathbf{g}$)

$$\epsilon(\pm\tfrac{3}{2})=\pm[1+(5/4)x^2+x(3+x^2)^{\frac{1}{2}}]^{\frac{1}{2}},$$
$$\epsilon(\pm\tfrac{1}{2})=\pm[1+(5/4)x^2-x(3+x^2)^{\frac{1}{2}}]^{\frac{1}{2}}.$$

[13] Standberg, Davis, Faughnan, Kyhl, and Wolga, Phys. Rev. **109**, 1988 (1958). See also Strandberg, Davis, and Kyhl, Fifth International Symposium on Low-Temperature Physics and Chemistry, Madison, Wisconsin (August 30, 1957).

[14] A. A. Manenkov and A. M. Prokhorov, Soviet Phys. JETP **1**, 611 (1955); M. M. Zaripov and Iv. Ia. Shamonin, Soviet Phys. JETP **3**, 171 (1956); J. E. Geusic, Phys. Rev. **102**, 1252 (1956).

[15] J. E. Geusic, Phys. Rev. **102**, 1252 (1956).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

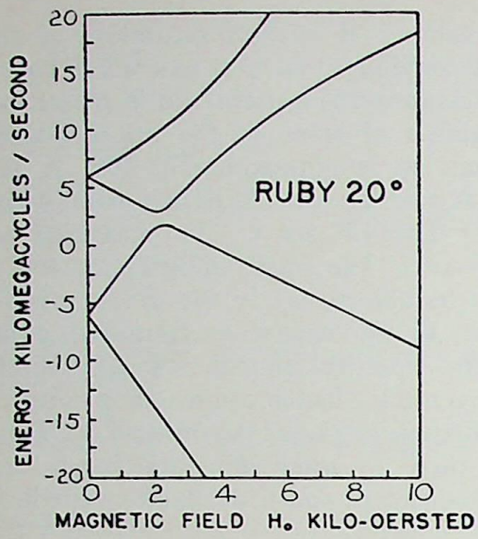


FIG. 10. Energy levels of Cr^{+++} in ruby *vs* applied magnetic field for H_o at 20° to c axis of the crystal.

Curves of energy levels as calculated with assistance of Dr. Schulz-DuBois are shown in Figs. 9–13. The four-level device using ruby has already been discussed. We are indebted to Professor Townes for the name "push-pull maser" for the four-level device of Makhov, Kikuchi, Lambe, and Terhune.‡

B. Characteristics of Some Ruby Masers

A ruby three-level maser has been developed by Alsop, Giordmaine, and Townes.[16] A large filling factor, approaching 0.9, is used. The "voltage gain" band-width product approaches 100 Mc, and the band width is approximately 5 Mc. The magnetic Q is about 400. The volume of the active material is about $\frac{1}{3}$ cm³. A rectangular cavity is used, operating in the TE_{011} mode for the signal frequency and the TE_{012} mode for the saturation microwave field. The small cavity volume results from the high dielectric constant of ruby. About 30 mw of pumping power are required. The pump power is coupled to the wave guide by a remotely gear-controlled probe, and the signal power is coupled in by means of an iris. Liquid helium cooling with helium maintained at low pressure provided a bath temperature of 1.4°K.

Morris, Kyhl, and Strandberg[17] have described the ruby maser illustrated in Fig. 14. All four levels are employed. The chromium concentration is about 0.01%. Levels 1–3 and 2–4 are saturated at 23 kMc/sec. Levels 2–3 are employed for amplification. The device is tunable over the range 8400–9700 Mc/sec, and has the new feature that cavity resonance is not needed at the saturation frequency if a saturation power of 100 mw is employed. The fringing fields near the coupling hole allow enough coupling to saturate the crystal at 23 kMc/sec.

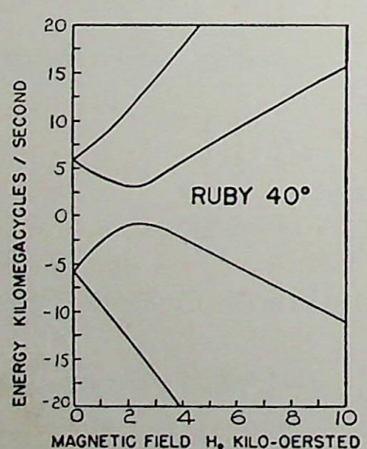


FIG. 11. Energy levels of Cr^{+++} in ruby *vs* applied magnetic field for H_o making an angle of 40° with the c axis of the crystal.

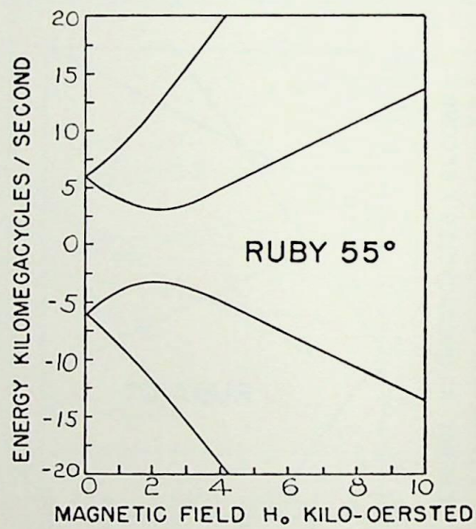


FIG. 12. Energy levels of Cr^{+++} in ruby *vs* applied magnetic field for H_o making an angle of 55° with the c axis of the crystal.

Design of masers is facilitated by the extensive tables of energy levels and transition probabilities now being prepared by W. S. Chang and A. E. Siegman of the Stanford Electronics Laboratories.[18]

IV. SUMMARY OF OTHER METHODS FOR OBTAINING MASER ACTION

Bloch[19] showed, in 1946, that inversion of level populations can be obtained by "adiabatic rapid pas-

‡ *Note added in proof.*—The paramagnetic resonance absorption for Cr^{+++} ions in emerald has been reported by Geusic, Peter, and Schulz-DuBois [Bull. Am. Phys. Soc. Ser. II, 4, 21 (1959)]. The spin Hamiltonian is

$$\mathcal{H}=\beta_m[g_{11}H_zS_z+g_1(H_xS_x+H_yS_y)]+D[S_z^2-\tfrac{1}{3}S(S+1)],$$

$2D=-53.6$ kmc, $g_{11}=1.973\pm0.002$, $g_1=1.97\pm0.01$.

[16] Alsop, Giordmaine, Mayer, and Townes, Astron. J. **63**, 301 (1958).

[17] Morris, Kyhl, and Strandberg, Proc. Inst. Radio Engrs. **47**, 80 (1959).

[18] W. S. Chang and A. E. Siegman, Stanford Electronics Lab. Tech. Rept. 156–1 (May 16, 1958). Their machine calculations are presented in the Appendix to this paper. Note carefully that their calculations were done using positive D. To use their curves and data for the correct negative sign of D, change the sign of their energies and regard their level labeled number four as having the lowest energy. Note their comments at the end of the Appendix.

[19] F. Bloch, Phys. Rev. **70**, 460 (1946). See also R. K. Wangsness and F. Bloch, Phys. Rev. **89**, 728 (1953).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

sage" through resonance. We have again a system of spins in a "constant" magnetic field H_o. If a microwave magnetic field is applied at right angles to the "constant" magnetic field the spin system will precess about H_o. Suppose that H_o is smaller than the value required for resonance at the microwave angular frequency ω. If H_o is steady for a long time, and is then suddenly increased through resonance and beyond, the spin system is turned over. The magnetization vector will be antiparallel to H_o until equilibrium is restored by the spin-lattice relaxation mechanism. The passage through resonance must be adiabatic, but rapid enough so that the sweep occurs in a time short compared with the spin-lattice relaxation time. While the magnetization is antiparallel to the field there are more moments antiparallel than parallel. This means more particles in excited states than in the ground state. Such a system will therefore amplify.

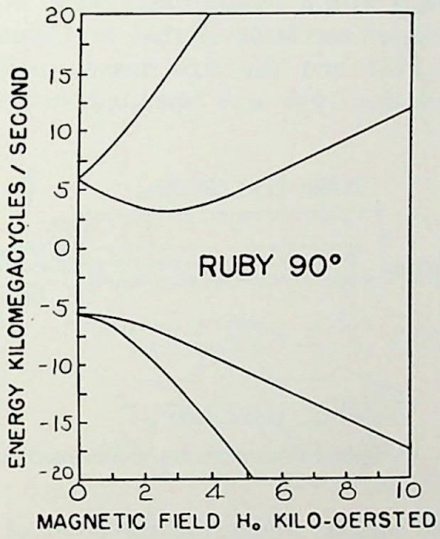


FIG. 13. Energy levels of Cr^{+++} in ruby *vs* applied magnetic field for H_o making an angle of 90° with the c axis of the crystal.

We assume the magnetic field H_o to be in the z direction. It is simpler to discuss the case where a circularly polarized rf field is applied, with its magnetic field in the x–y plane. The rf field is given by

$$H_x = H_1 \cos\omega t, \quad H_y = \mp H_1 \sin\omega t. \tag{36}$$

The $-$ sign refers to a positive gyromagnetic ratio and the $+$ sign to a negative one. The magnetization vector then satisfies the following equations.

$$\dot{M}_x - \gamma(M_y H_z - M_z H_y) + (M_x/T_2) = 0, \tag{37}$$

$$\dot{M}_y - \gamma(M_z H_x - M_x H_z) + (M_y/T_2) = 0, \tag{38}$$

$$\dot{M}_z - \gamma(M_x H_y - M_y H_x) + (M_z/T_1) = M_o/T_1. \tag{39}$$

γ is the gyromagnetic ratio, T_1 is the longitudinal (spin-lattice) relaxation time, T_2 is the transverse (spin-spin) relaxation time, and M_o is the value of the magnetization in the absence of a radio-frequency field.

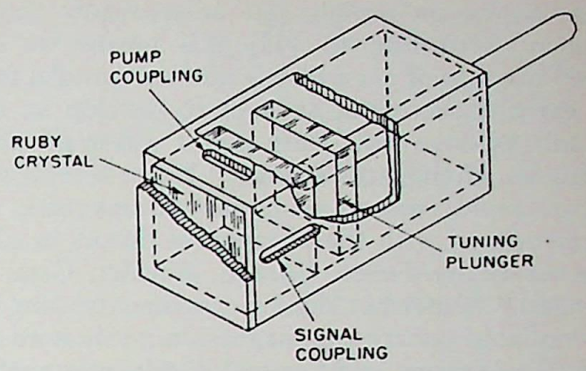


FIG. 14. Ruby maser of Morris, Kyhl, and Strandberg.

A quantity $\delta(t)$ is defined by the relation

$$\delta(t) = \frac{H_o(t) - (\omega/|\gamma|)}{H_1}. \tag{40}$$

$\delta(t)$ is zero at the resonant field $H_o = \omega/|\gamma|$. The magnetic field H_o is assumed to vary at a slow enough rate so that

$$|d\delta/dt| \ll |\gamma H_1|. \tag{41}$$

This is a statement of the adiabatic criterion. Subject to this condition Bloch gives the following solution of (37), (38), and (39):

$$M_x = \frac{M\cos\omega t}{(1+\delta^2)^{\frac{1}{2}}}, \tag{42}$$

$$M_y = \mp\frac{M\sin\omega t}{(1+\delta^2)^{\frac{1}{2}}}, \tag{43}$$

$$M_z = M\delta/(1+\delta^2)^{\frac{1}{2}}, \tag{44}$$

$$M(t) = \frac{1}{T_1}\int_{-\infty}^{t} \frac{\delta(t')e^{-[\theta(t)-\theta(t')]}M_o(t')}{[1+\delta^2(t')]^{\frac{1}{2}}}dt', \tag{45}$$

$$\theta(t) - \theta(t') = \frac{1}{T_1}\int_{t'}^{t}\left[\frac{\delta^2(t'') + (T_1/T_2)}{1+\delta^2(t'')}\right]dt''. \tag{46}$$

In order to discuss adiabatic rapid passage we assume that $\delta^2(t'')$ has been constant for a long time and then at time t_o is quickly increased through resonance, without violating (41). We have, for $\delta^2 \gg T_1/T_2$,

$$\theta(t) - \theta(t') = (t-t')/T_1. \tag{47}$$

The part of (45) resulting from the rapid change of δ is negligible. For the rest, since δ is large we have, from (45)

$$M(t) = \pm M(t_o). \tag{48}$$

Here the $+$ sign refers to the situation where δ was positive up to t_o (H_o larger than the resonant value, initially), and the $-$ sign refers to the situation where δ was negative up to t_o (H_o smaller than the resonant value, initially, then increased through resonance). Thus the sign of $M(t)$ does not change, but the sign of δ

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

changes as we go rapidly but adiabatically through resonance. According to (44) this means we have changed the sign of M_z which is now antiparallel to H_o.

An early unsuccessful attempt to develop an adiabatic fast passage solid-state maser was reported by Combrisson, Honig, and Townes.[20] Maser action due to adiabatic rapid passage using the paramagnetic electrons associated with the phosphorus donors in silicon was observed by Feher, Gordon, Buehler, Gere, and Thurmond.[21] Somewhat similar experiments using neutron irradiated quartz and magnesium oxide were done by Chester, Wagner, and Castle.[22] With a quartz sample containing $\sim 10^{18}$ spins the inverted state persisted for 2 msec at 4.2°K. A value of 5×10^6 sec^{-1} was obtained for the "voltage" gain-band-width product for gains between 8 and 21 db. This work was done at 9 kMc/sec with microwave powers for inversion of about 0.5 w, in 50–100 μsec pulses. A modulation structure of the pulsed power emitted from the cavity as the magnetic field is swept through resonance is observed. Senitzky[23] has suggested that this amplitude modulation of the power from the resonant cavity is due to the periodic transfer of energy between the cavity electromagnetic field and the spin system.

Purcell and Pound[24] were able to obtain inverted level populations in a magnetic resonance experiment using a single crystal of LiF which had a very long relaxation time. They removed their crystal from the strong field and inverted its spin system by means of other rapidly varying magnetic fields. When the crystal was reinserted in the strong field they were able to observe the return of the magnetization to its equilibrium value, from its "negative-temperature" state.

Weber studied the following method[1] in 1951. Consider a symmetric top molecule in an externally applied electrostatic field. The Stark effect has linear and quadratic terms. The linear effect is the dominant one if the field is not too strong. The energy levels are given by the formula

$$E_{JKM_j}^{(1)}=\frac{-\boldsymbol{\mu}\cdot\mathbf{E}KM_j}{J(J+1)}. \tag{49}$$

$\boldsymbol{\mu}$ is the electric dipole moment, $\mathbf{E}$ is the Stark field, J, K, and M_j are the quantum numbers associated with the symmetrical top. If the electrostatic (Stark) field is applied normal to the microwave electric field the $\Delta M=\pm 1$ transition will be allowed, and the frequency for such a transition will be

$$\nu=\boldsymbol{\mu}\cdot\mathbf{E}K/J(J+1)h. \tag{50}$$

If the gas is in equilibrium it will absorb microwaves at this frequency. There are more electric dipole moments parallel to the field than antiparallel to it. If the electrostatic field is suddenly reversed we have a negative temperature, and more dipole moments antiparallel than parallel to the field, and the device will amplify during roughly one relaxation time. A pulsed oscillator may be constructed as shown in Fig. 15. Here we have a resonant cavity with Stark electrode. If a square wave is applied there will be a microwave pulse emitted each time the field reverses and the TE_{10} mode (microwave electric field *parallel* to the Stark electrode) will be excited.

No experiments of this type were done because calculations showed that it would be difficult to achieve a useful gain-band width product in any gas-type maser amplifier, and the use of a solid was suggested.

A maser oscillator using a gas is, however, a very useful frequency standard, and compares favorably in many respects with a cesium beam clock. Work on an ammonia maser oscillator started at Columbia University in 1951 and the first maser oscillator and amplifier of this type was operated successfully by

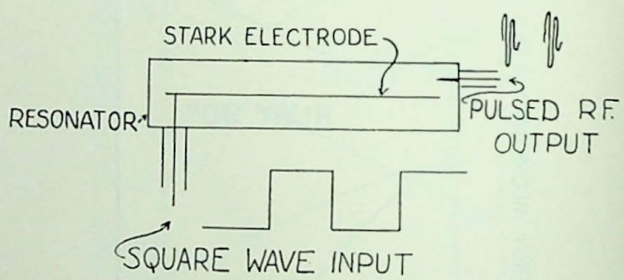


FIG. 15. Pulsed symmetric top maser oscillator.

Gordon and Zeiger[2] under the direction of Professor C. H. Townes in 1954. Similar work was done independently[3] by Basov and Prokhorov at the Lebedev Institute at about the same time. In addition to providing an exceptionally stable oscillator, the ammonia maser can also be employed as a very high resolution spectrometer. A sketch of the focuser and beam and cavity details is given in Fig. 16. The maser action is accomplished in the following way. The nitrogen atom in ammonia is in a potential which is symmetric on either side of the plane of the hydrogen atoms. In this kind of potential each harmonic oscillator level is split, and the ground state splitting provides microwave absorption lines of different frequencies for different rotational levels. In a Stark field the energy of the upper doublet level increases and that of the lower level decreases. The focuser of Fig. 16, provides an electric field with intensity approximately proportional to the displacement from the axis. If a molecular beam is now sent through the focuser the molecules in the lower doublet state drift outward to regions of lower energy and increasing field, and are removed. The center of the beam contains mostly higher energy particles. These

[20] Combrisson, Honig, and Townes, Compt. rend. **242**, 2451 (1956). Suggestions for a low-temperature adiabatic rapid passage maser were made independently by Strandberg in 1956.
[21] Feher, Gordon, Buehler, Gere, and Thurmond, Phys. Rev. **109**, 221 (1958).
[22] Chester, Wagner, and Castle, Phys. Rev. **110**, 281 (1958).
[23] I. R. Senitzky, Phys. Rev. Letters **1**, 167 (1958).
[24] E. M. Purcell and R. V. Pound, Phys. Rev. **81**, 279 (1951).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

enter the resonant cavity and undergo stimulated emission. In this case, then, a higher population in excited states is achieved by the simple expedient of removing ground state particles from the beam.

The condition for sustained oscillation of a resonant cavity with volume V which contains n excited atoms or molecules may be obtained by setting the emitted power equal to the power lost in the cavity walls.

$$nW_{21}h\nu \geq P_w.$$

W_{21} is the transition probability. We write W_{21} in terms of the appropriate squared matrix elements $|\mu|^2$ and the line width $\Delta\nu$, and P_w in terms of the quality factor Q to obtain (Gordon, Zeiger, and Townes, see reference 2)

$$n \gtrsim hV\Delta\nu/4\pi|\mu|^2Q.$$

The frequency range over which appreciable energy is distributed in a maser oscillator is given in terms of the

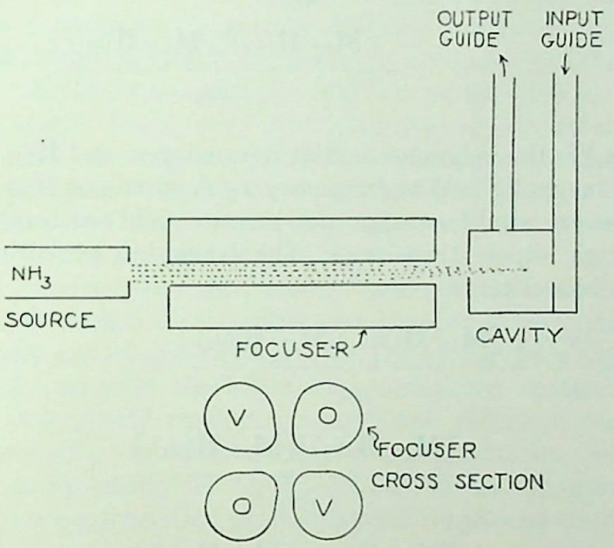


FIG. 16. Columbia ammonia maser.

total power output P, Boltzmann's constant, and the cavity wall temperature T as

$$\delta\nu = 4\pi kT(\Delta\nu)^2/P.$$

An upper state population ten times that of the lower state may be achieved in practice. The power output is approximately 10^{-9} w. An ammonia maser oscillator has been reported to have frequency stability over periods of one hour of the order of 1 part in 10^{10}. The stability is associated with the low noise property.

Another method of achieving inverted level populations in ammonia gas has been proposed by Dicke.[25] This uses a "hot grid" cell and again the quadratic Stark effect. A maser using this method of population inversion is being constructed by Dr. J. P. Wittke at RCA Laboratories.

Level population inversion can also be achieved by means of intense pulses of microwaves. For simplicity,

[25] J. P. Wittke, Proc. Inst. Radio Engrs. 45, 291 (1957).

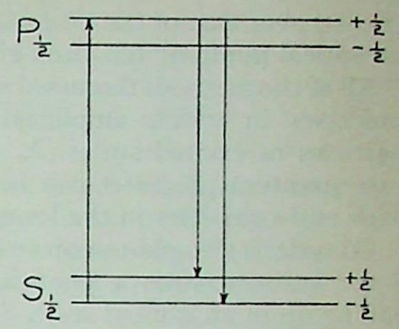


FIG. 17. Optical pumping.

consider, a two-level system with "permanent" electric dipole moment μ, driven at resonance[26] by a microwave field $E=E_o\sin(\omega t-\delta)$. The Hamiltonian is

$$\mathcal{H}=\mathcal{H}_o+\mathfrak{u}\cdot\mathbf{E}_o\sin(\omega t-\delta). \quad (51)$$

Let the wave functions with $E_o=0$ be ψ_1 and ψ_2.

If we use this in the time-dependent Schroedinger equation and neglect sum frequency terms, an exact solution of the resulting equations gives for the wave function,

$$\Psi=\left[\cos\left(\frac{|\mathfrak{u}_{12}\cdot\mathbf{E}_o|t}{2\hbar}\right)\right]\psi_1 e^{-i(E_1t/\hbar)} + \left[\sin\left(\frac{|\mathfrak{u}_{12}\cdot\mathbf{E}_o|t}{2\hbar}\right)\right]\psi_2 e^{-iE_2t/\hbar}. \quad (52)$$

Here $\mathfrak{u}_{12}$ is the dipole matrix element connecting the states 1 and 2. From this it follows that if the system is in the state ψ_1 at $t=0$, then at $t=(\pi\hbar)/(|\mathfrak{u}_{12}\cdot\mathbf{E}_o|$, the microwave pulse has driven the system into state 2. This treatment is valid only for times short compared with the relaxation time since the effect of collisions is ignored.

Thus if we have thermal equilibrium at $t=0$ and apply a strong rf pulse of the correct length we can obtain inversion of level populations. This effect has been demonstrated by Dicke and Romer,[26] and by Kyhl, Standberg, Collins, and Park.[27]

The methods of optical pumping can also be employed[28] to achieve maser action. In Fig. 17 we have a system with a $^2S_{\frac{1}{2}}$ ground state and a $^2P_{\frac{1}{2}}$ excited state, in a magnetic field. In equilibrium almost all of the particles will be in the split S state. The $m=-\frac{1}{2}S$ state will have slightly more particles in it than the $m=+\frac{1}{2}$ state. Right circularly polarized optical resonance radiation will induce transitions with the selection rule $\Delta m=+1$. The excited particles can now spontaneously emit, either with $\Delta m=-1$, or 0. However this process only removes particles from the $m=-\frac{1}{2}S$ state and returns them to *both* the $m=-\frac{1}{2}$ and $m=+\frac{1}{2}$ states. Therefore the population of the $^2S_{+\frac{1}{2}}$ state will be in-

[26] R. H. Dicke and R. H. Romer, Rev. Sci. Instr. 26, 915 (1955)

[27] Kyhl, Strandberg, Collins, and Park, Signal Corps Symposium on Solid State Masers, Fort Monmouth, New Jersey (June 12–13, 1958).

[28] W. H. Culver, Science 126 810 (1957).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

creased over that of the $^2S_{-\frac{1}{2}}$ state. A review of the work on optical pumping has been given by Kastler.[29]

All of the methods discussed so far use single quantum processes in which amplification results from more particles in excited states. A. Javan[30] has shown that two quantum processes can be used for amplification, with more particles in the lower state (Fig. 18).

Consider a two photon process in which a particle in a lower state absorbs a photon, then emits a photon, ending up in an excited state. Suppose the frequency of absorption corresponds to that of an intense (pump) oscillator and the emission is stimulated by a signal which it is desired to amplify. In this case we are obtaining emission in a particle transition from a state of lower energy to a state of higher energy. If all particles are in the ground state, amplification will result. Let the transition probability be W_{12} and let n_1 particles be in the ground state. The emitted power is $W_{12}n_1h\nu$. If the excited state has n_2 particles the reverse process may occur in which a photon of amplification frequency ν is absorbed and a pump frequency photon is emitted. The net power emitted will be $W_{12}h\nu(n_1-n_2)$, which is positive at positive temperatures.

For a *one* quantum process W_{12} may be written as

$$W_{12}=4\pi^2\tau|H_{12}'|^2/h^2. \tag{53}$$

Here τ is the relaxation time and H' is the interaction matrix element. For a two quantum process H' is replaced by

$$\sum_i H_{1i}'H_{i2}'/(E_1-E_i). \tag{54}$$

The summation is over all intermediate states of the particle and the quantized electromagnetic field. In Raman spectroscopy the intermediate states are usually quantum states which are different from the initial and final states, because the diagonal interaction matrix elements ordinarily vanish. If the diagonal matrix elements do not vanish in either the initial or final state, no intermediate state is necessary. For example we may have a diagonal magnetic dipole matrix element in a product of the following type

$$\langle-\tfrac{1}{2},N_1,N_2|H'|-\tfrac{1}{2},N_1+1,N_2\rangle \times\langle-\tfrac{1}{2},N_1+1,N_2|H'|N_2-1,N_1+1,+\tfrac{1}{2}\rangle. \tag{55}$$

Here we use the nomenclature of quantum electrodynamics. The first bracket involves a matrix element between states in which the particle spin is $-\frac{1}{2}$, and the electromagnetic field oscillators of frequency ν_1 and ν_2 have N_1 and N_2 quanta, respectively, to a state with particle spin unchanged, N_1+1 quanta for the oscillator with frequency ν_1, and N_2 quanta for the oscillator with frequency ν_2. The second bracket then involves a final state in which the spin becomes $+\frac{1}{2}$ and the field oscillator of frequency ν_2 has lost a quantum. Three additional matrix elements, similar to (55) may be written. One of these corresponds to an intermediate state of the particle the same as the initial state but an absorption first of frequency ν_2 followed by emission of frequency ν_1. The other two matrix elements involve an intermediate state of the particle which is the same as the final state, with both possibilities for the order of absorption and emission. Some of these matrix elements may be small in comparison with the others, depending on circumstances of the experimental arrangement. The summation (54) may be written in terms of the (magnetic) dipole matrix elements M_{11} and M_{12} as

$$-\frac{[\mathbf{M}_{11}\cdot\mathbf{H}(\nu_1)][\mathbf{M}_{12}\cdot\mathbf{H}(\nu_2)]}{h\nu_1} +\frac{[\mathbf{M}_{12}\cdot\mathbf{H}(\nu_2)][\mathbf{M}_{22}\cdot\mathbf{H}(\nu_1)]}{h\nu_1} +\frac{[\mathbf{M}_{11}\cdot\mathbf{H}(\nu_2)][\mathbf{M}_{12}\cdot\mathbf{H}(\nu_1)]}{h\nu_2} -\frac{[\mathbf{M}_{12}\cdot\mathbf{H}(\nu_1)][\mathbf{M}_{22}\cdot\mathbf{H}(\nu_2)]}{h\nu_2}. \tag{56}$$

$\mathbf{H}(\nu_1)$ is the magnetic field at frequency ν_1 and $\mathbf{H}(\nu_2)$ is the magnetic field at frequency ν_2. A corresponding expression would contain the electric field vectors for electric dipole transitions. The transition probability can be written as

$$W=\frac{4\pi^2\tau}{h^2}\left|-\frac{[\mathbf{M}_{11}\cdot\mathbf{H}(\nu_1)][\mathbf{M}_{12}\cdot\mathbf{H}(\nu_2)]}{h\nu_1} +\frac{[\mathbf{M}_{12}\cdot\mathbf{H}(\nu_2)][\mathbf{M}_{22}\cdot\mathbf{H}(\nu_1)]}{h\nu_1} +\frac{[\mathbf{M}_{11}\cdot\mathbf{H}(\nu_2)][\mathbf{M}_{12}\cdot\mathbf{H}(\nu_1)]}{h\nu_2} -\frac{[\mathbf{M}_{12}\cdot\mathbf{H}(\nu_1)][\mathbf{M}_{22}\cdot\mathbf{H}(\nu_2)]}{h\nu_2}\right|^2. \tag{57}$$

The net power emitted by a two-level system which has n_1 particles in the lower state and n_2 particles in the upper state may then be written as

$$P=\frac{4\pi^2(n_1-n_2)\tau\nu_1}{h}\left|-\frac{[\mathbf{M}_{11}\cdot\mathbf{H}(\nu_1)][\mathbf{M}_{12}\cdot\mathbf{H}(\nu_2)]}{h\nu_1} +\frac{[\mathbf{M}_{12}\cdot\mathbf{H}(\nu_2)][\mathbf{M}_{22}\cdot\mathbf{H}(\nu_1)]}{h\nu_1} +\frac{[\mathbf{M}_{11}\cdot\mathbf{H}(\nu_2)][\mathbf{M}_{12}\cdot\mathbf{H}(\nu_1)]}{h\nu_2} -\frac{[\mathbf{M}_{12}\cdot\mathbf{H}(\nu_1)][\mathbf{M}_{22}\cdot\mathbf{H}(\nu_2)]}{h\nu_2}\right|^2. \tag{58}$$

[29] A. Kastler, J. Opt. Soc. Am. 47, 460 (1957).
[30] A. Javan, Bull. Am. Phys. Soc. Ser. II, 3, 213 (1958).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

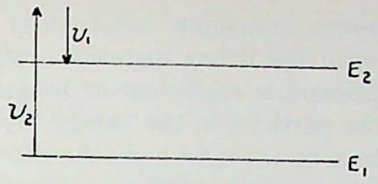


FIG. 18. Raman maser.

V. NOISE PERFORMANCE

Prior to 1950, methods had been developed for generation, amplification, and detection of signals in the electromagnetic spectrum from zero frequency to the mm wave region. At frequencies below 10 mc/sec the noise performance is already better than is ordinarily required for communications in the presence of atmospheric and other external noise. Indeed, a well designed vacuum tube amplifier in the vicinity of 1 mc/sec is capable of noise performance corresponding to an internal temperature of about 25°K. Microwave amplifiers, on the other hand, did not have good low-noise performance, mainly because of electron stream fluctuations originating at a hot cathode. Inasmuch as this type of noise must be absent in a maser, it was suggested that molecular beam maser type amplifiers would have very little noise.[31] It was very gratifying that the important sources of low-temperature solid state maser noise which were understood in 1956, are sufficiently small to make possible sensitivities which may approach detection of single microwave photons.

We consider first the inherent spontaneous emission noise of a maser, disregarding circuit noise, at spin temperature approaching minus 0°K. The earlier discussion did not consider the effect of spontaneous emission. This is a purely random process[32] and therefore contributes noise. In an ensemble of charged particles, each having two energy levels, those which are in the upper state have a transition probability for transitions to the lower state of the form

$$W_{21}=f(\nu,\mu^2\cdots)[N+1], \tag{59}$$

in consequence of the interaction with the electromagnetic field.§

Here f is a function of frequency and squared matrix elements and N is the number of quanta per radiation oscillator. The one which adds to N gives the effect of spontaneous emission. Let us calculate an equivalent temperature for spontaneous emission. In thermal equilibrium at temperature T, according to the Bose-Einstein statistics, N would be

$$\frac{1}{e^{h\nu/kT}-1}. \tag{60}$$

The positive temperature which black surroundings would need, to emit input noise equivalent to the spontaneous emission, will be denoted by T_E. From (59) we see that this (equivalent) T_E corresponds to an equivalent $N=1$.

Expressions (60) and (59) allow us to write

$$1=e^{h\nu/kT_E}-1, \tag{61}$$

giving

$$T_E=h\nu/k\,\ln_e 2. \tag{62}$$

Suppose we ask the question: how many microwave photons will double the randomly fluctuating output of a maser whose spin temperature approaches -0°K. Let the averaging time of the receiver be τ and let the equivalent input noise energy during this interval be U_n. This is given by the product of the Nyquist formula noise power for T_E and the averaging time τ.

$$U_n=\left[\frac{h\nu}{e^{h\nu/kT_E}-1}\right]\Delta\nu\tau. \tag{63}$$

Here $\Delta\nu$ is the band width. $\Delta\nu\tau\approx 1$. Employing expression (61) gives

$$U_n=h\nu. \tag{64}$$

The equivalent input (noise) energy over the averaging time is that of one photon. If one microwave photon is incident during this interval it will double the average noise output and therefore we are justified in saying that a maser with no circuit noise and spin temperature approaching -0°K may detect single microwave photons, and a flux of $\Delta\nu$ photons per sec, with reasonably large probability.

The foregoing analysis gives exactly what would be predicted by the Nyquist formula for *any* resistance at a temperature -0°K. The Nyquist formula noise energy (delivered to a "matched" load) over the averaging time τ is given by

$$U_n=\frac{h\nu\Delta\nu\tau}{|e^{h\nu/kT}-1|}. \tag{63a}$$

In the limit as $T\rightarrow -0°$ (63a) becomes

$$U_n=h\nu\Delta\nu\tau\approx h\nu, \tag{64a}$$

in agreement with (64). This makes it clear that the residual noise in a low-temperature maser is of the same type as discussed by Johnson and Nyquist[33] long ago.

If the spin temperature is not close to minus 0°K (64a) becomes

$$U_n=\frac{h\nu\Delta\nu\tau}{|e^{h\nu/kT_{\text{spin}}}-1|}. \tag{64b}$$

[31] The Columbia work (Gordon, Zeiger, and Townes, reference 2) showed that good noise performance could be obtained with a molecular beam maser. Spontaneous emission from the molecules was not discussed. For a room temperature device this is small compared with input noise.

[32] It has been shown [J. Weber, Phys. Rev. 94, 215 (1954)] that this type of noise is also present in conventional free electron amplifiers.

§ *Note added in proof.*—Expression (59) is given in Chapter V, *The Quantum Theory of Radiation*, by W. Heitler (Oxford University Press, London, 1954), third edition. It follows from the quantization of the Maxwell field. Each degree of freedom is like a harmonic oscillator and the $N+1$ arises from the squared harmonic oscillator matrix elements for downward transitions.

[33] J. B. Johnson, Phys. Rev. 32, 47 (1928); H. Nyquist, Phys. Rev. 32, 110 (1928).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

For the case $(h\nu)/(kT)\ll 1$, the number of photons required to double the output is therefore

$$|kT_{\text{spin}}/h\nu|. \tag{65}$$

Javan has remarked that the noise performance of a Raman type maser will be essentially the same as that of devices which employ greater populations of particles in excited states. This follows because the photon of frequency ν_1 (Fig. 18) can be spontaneously emitted in the absence of a signal, provided that intense "pump" radiation of frequency ν_2 is present.

A complete system will also have noise due to emission from circuit elements and the transmission system. A quantity which has been used to describe the performance of radio receivers is the noise figure, denoted by the symbol F. F is defined as the ratio of the total noise power output of the receiver to that part of the noise power output due to the source at the input of the receiver. An equivalent definition is that it is the quotient of the signal to noise power ratio at the input divided by the signal to noise power ratio at the output.

We now calculate the noise figure of a complete maser receiving system, considering first a traveling wave maser. Our approach is similar to that employed by Strandberg,[34] for two-level systems.[35] We consider three-level systems, taking effects of saturation[36] into account. Let $R_\nu d\nu$ be the number of modes per unit volume in a range $d\nu$ which can propagate, A be the cross-sectional area, and V_G the group velocity. The number of particles per unit volume in the states with energies E_1, E_2, and E_3 will be denoted by n_1, n_2, and n_3. Let β be the power gain per unit length. A quantity K is defined by the following relation

$$\beta = K(n_2 - n_1). \tag{66}$$

Let N be the average number of quanta per mode in the vicinity of the operating frequency ν_{21} and let p_ν be the energy per mode. The change in power as a consequence of energy exchanged in a differential length leads to

$$\begin{aligned} &V_G A R_\nu d\nu (dp_\nu/dx) \\ &\quad = V_G A R_\nu d\nu [Kn_2 h\nu(N+1) - Kn_1 h\nu N - \alpha_c N h\nu \\ &\qquad + \alpha_c p_\nu(T_c) - \alpha_{13} N h\nu + \alpha_{13} p_\nu(T_{13})]. \end{aligned} \tag{67}$$

In expression (67) $p_\nu(T_c)$ is the average energy per mode in equilibrium at the temperature of the waveguide walls T_c, and from statistical mechanics $p_\nu(T) = h\nu[e^{h\nu/kT}-1]^{-1}$, $p_\nu(T_{13})$ is the average energy per mode at the temperature T_{13} associated with particles in the energy levels with energies E_1 and E_3, α_c is the power absorption coefficient of the transmission system when the maser material is absent and $\alpha_{13}(\nu_{21})$ is the absorption coefficient at frequency ν_{21} associated with the particles in the energy levels E_1 and E_3. The first term on the right gives the effect of both stimulated and spontaneous emission in inducing transitions out of the state with energy E_2, the second term gives the absorption by particles in the ground state, the third term gives the effect of absorption by conducting walls or loss other than the active solid, the fourth term gives the emission from the walls, the fifth and sixth terms represent the effects of absorption and emission from particles undergoing transitions between the first and third states. We define the temperatures T_{12} and T_{13} by

$$n_1/n_2 = e^{h\nu_{21}/kT_{12}}, \quad n_1/n_3 = e^{h\nu_{31}/kT_{13}}. \tag{68}$$

Integrating expression (67) gives

$$(p_\nu)_{\text{out}} = (p_\nu)_{\text{in}} g^2 + \left[\frac{\beta p_\nu(T_{12}) - \alpha_c p_\nu(T_c) - \alpha_{13} p_\nu(T_{13})}{\beta - \alpha_c - \alpha_{13}}\right] \times (1-g^2). \tag{69}$$

Here g^2 is the power gain defined earlier. This tells us that the temperature of the environment of the maser enters in a subtle way, through the quantity $\alpha_c p_\nu(T_c)$. If the absorption coefficient of the environment, α_c, is small, then little noise is contributed even if T_c is not close to absolute zero. This is a consequence of Kirchhoff's law. Let the source temperature be T_s, the load temperature be T_L, and let the transmission line to the maser have a power loss factor t and a temperature T_t. Thermodynamic considerations enable us to express the output noise emission of the transmission line per cycle in terms of the input noise emission (per cycle) of the source, as

$$(p_\nu)_{\text{line}} = t p_\nu(T_s) + (1-t) p_\nu(T_t). \tag{70}$$

Expression (70) is $(p_\nu)_{\text{in}}$ for the maser.

We can now use (69), (70), and the previous definition of noise figure to write

$$\begin{aligned} F &= \frac{(p_\nu)_{\text{out}} + p_\nu(T_L)}{g^2 t p_\nu(T_s)} \\ &= 1 + \frac{1}{t p_\nu(T_s)} \left\{ (1-t) p_\nu(T_t) + (1-g^{-2}) \right. \\ &\quad \left. \times \left[\frac{\alpha_c p_\nu(T_c) + \alpha_{13} p_\nu(T_{13}) - \beta p_\nu(T_{12})}{\beta - \alpha_c - \alpha_{13}}\right] + \frac{p_\nu(T_L)}{g^2} \right\}. \end{aligned} \tag{71}$$

It was once thought that the saturation field would make a significant contribution to the noise because the temperature T_{13} associated with the (saturated) 1-3 system tends towards infinity. This is not the case; only a negligible contribution results, in the following way.[36]

[34] M. W. P. Strandberg, Phys. Rev. **106**, 617 (1957).

[35] Noise in maser type amplifiers has been discussed by Muller [Phys. Rev. **106**, 8 (1957)], Pound [Ann. Phys. **1**, 24 (1957)] and Shimoda, Takahasi, and Townes [J. Phys. Soc. Japan **12**, 686 (1957)]. The role of spontaneous emission was emphasized by R. H. Dicke at the Symposium on Amplification by Atomic and Molecular Resonance, Asbury Park, New Jersey (March 1, 1956).

[36] J. Weber, Phys. Rev. **108**, 537 (1957).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

The quantity $\alpha_{13}(\nu_{21})$ may be written

$$\alpha_{13}(\nu_{21})=\frac{\chi n_1|\mu_{13}|^2}{kT_{13}}\left[\frac{\tau_{13}^{-1}}{4\pi^2(\nu_{21}-\nu_{31})^2+\tau_{13}^{-2}}\right], \quad (72)$$

where χ is a constant, and τ_{13} is the appropriate relaxation time.

β may be written in terms of χ and the 1–2 system relaxation time τ_{12}, as

$$\beta=\frac{\chi n_2|\mu_{12}|^2\tau_{12}}{|p_\nu(T_{12})|}. \quad (73)$$

The ratio of the terms involving α_{13} and β in the noise figure formula (71) is

$$\frac{\alpha_{13}p_\nu(T_{13})}{|\beta p_\nu(T_{12})|}=\frac{n_1|\mu_{13}|^2}{n_2|\mu_{12}|^2}\left[\frac{(\tau_{13}\tau_{12})^{-1}}{4\pi^2(\nu_{21}-\nu_{31})^2+\tau_{13}^{-2}}\right]. \quad (74)$$

For a typical amplifier in the microwave region, this gives a number of order 10^{-3}. Physically this is because a high-saturation temperature means a small absorption coefficient for the 1–3 system, consequently small noise emission.

In order to extend the previous results to a cavity type of maser we note that we have one mode with effective power width $(\pi/2)\Delta\nu$. $\Delta\nu$ is the band width over which the power exceeds one half of the maximum (exact resonance) power. The $\pi/2$ takes account of those parts of the response outside of the region where the power exceeds half the power at resonance. In order to calculate the noise we again employ detailed balancing on a power per cycle basis, making use of the definitions of the various Q's employed earlier. The noise power output per cycle is

$$P_{\text{noise}}=(g+1)^2\Big[lp_\nu(T_s)+(1-l)p_\nu(T_l)+\frac{Q_e}{Q_o}\Big\{p_\nu(T_c)-\frac{1}{1+\gamma}(p_\nu(T_{12})+\gamma p_\nu(T_{13}))\Big\}-\frac{g-1}{g+1}\Big(\frac{p_\nu(T_{12})+\gamma p_\nu(T_{13})}{\gamma+1}\Big)\Big]. \quad (75)$$

The noise figure is then given by

$$F=\frac{(g+1)^2}{g^2lp_\nu(T_s)}\Big[(1-l)p_\nu(T_l)+lp_\nu(T_s)+\frac{Q_e}{Q_o}\Big\{p_\nu(T_c)-\frac{1}{1+\gamma}(p_\nu(T_{12})+\gamma p_\nu(T_{13}))\Big\}-\frac{(g-1)}{(g+1)}\Big(\frac{p_\nu(T_{12})+\gamma p_\nu(T_{13})}{1+\gamma}\Big)\Big]+\frac{g^{-2}p_\nu(T_L)}{lp_\nu(T_s)}. \quad (76)$$

Here the quantity γ is defined by

$$\gamma=\frac{Q_m}{Q_{13}}=\frac{-\alpha_{13}(\nu_{21},T_{13})}{\beta}, \quad (77)$$

with Q_{13} the Q associated with energy absorption by the 1–3 system. The somewhat unusual appearance of (77) with the factor $(1+g)^2$ results from the fact that g^2 is the power gain at resonance. Different parts of the cavity response contribute noise, but with different effective gain.

The term noise temperature is often used, rather than noise figure. This may be defined in terms of the noise figure as

$$\left.\begin{aligned}&[e^{h\nu/kT_n}-1]^{-1}=(F-1)[e^{h\nu/kT_s}-1]^{-1}\\ &T_n\approx(F-1)T_s, \quad \text{for} \quad \frac{h\nu}{kT}\ll 1\end{aligned}\right\}. \quad (78)$$

Here T_n is the noise temperature, and T_s is the source temperature.

McWhorter and Arams[37] have measured the noise temperature of a complete solid state maser system and found it to be $20\pm5°$K.

The use of the concept of noise figure or noise temperature was very meaningful for the older type of microwave amplifiers. We propose a method of describing the noise performance which seems better for very quiet amplifiers. The number of photons received over the receiver averaging time which will double the noise output can be employed as a "noise number" to specify the performance of a low noise receiver. We showed earlier that the minimum noise number of a maser is one. The connection between noise number N_n and noise temperature T_n is

$$N_n=[e^{h\nu/kT_n}-1]^{-1}+[e^{h\nu/kT_s}-1]^{-1}. \quad (79)$$

VI. MICROWAVE PHOTON COUNTERS

The maser makes use of stimulated emission. If *all* of the particles were in the upper state of a two-level maser, it could amplify without absorbing any of the incident photons. However, under these conditions the noise number is one, so that at least one incident photon is necessary in order that its randomly fluctuating output be doubled over the receiver averaging time. The incidence of one photon could be interpreted as a spontaneous fluctuation. In an earlier paper[36] we remarked that a maser is a voltage amplifier, not a power amplifier, and that all voltage amplifiers have spontaneous emission noise. Detectors such as nuclear counters must absorb energy in order to operate. However, unlike a maser, the internal fluctuations of a counter or power amplifier can be made arbitrarily

[37] McWhorter and Arams, Lincoln Laboratory, MIT, Rept. M 37–22. See also Alsop, Giordmaine, Townes, and Wang, Phys. Rev. **107**, 1450 (1957); J. C. Helmer, Phys. Rev. **107**, 902 (1957).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

small. Such a device has essentially zero output until it detects a photon. It was proposed[36] that power amplifiers and detectors be developed which employ particles initially in their ground states. Methods for doing this are now being studied here. Consider a three-level system, in which the frequency ν_{21} (Fig. 19) for transitions from the first to the second state, is in the optical region. A lamp illuminates with intense light of frequency ν_{21} which can be linearly polarized. Then transitions from state 1 to state 2 are allowed, since $\Delta m=0$. Excited particles in state 2 can spontaneously emit only linearly polarized light. We employ a detector which counts circularly polarized optical frequency photons, which can be emitted only in certain directions. In order to detect microwave photons we arrange to have them circularly polarized. Then a microwave photon can be absorbed in a transition to state 3 or 4, since $\Delta m=1$. Excited particles in states 3 or 4 can now spontaneously emit circularly polarized photons which will be counted. Thus if no microwave photons are present we have particles in states 1 and 2 and only linearly polarized photons. The device has no output. As soon as a microwave (or infrared) photon is absorbed we have a particle in state 3 or 4, and when this particle emits, we have a circularly polarized photon, which is counted. This method is similar to that which has been employed in optical pumping experiments. The transition probability for this type of third-order process can be calculated in the following way. Let the wave function for atom and electromagnetic field be $\Phi(t)$.

$$\Phi(t)=\sum a_i(t)\Psi_i.$$

The Ψ_i are unperturbed wave functions of particle and electromagnetic field. Let the ground state be denoted by the subscript M. Let $a_M=1$ at $t=0$, and $a_j=0$ for $j\neq M$, at $t=0$. Let H' be the interaction part of the Hamiltonian. We quantize the electromagnetic field, H' is then not time dependent. Our process is one involving three photons. One is absorbed from the lamp, one is absorbed from the microwave field, and one is then spontaneously emitted. The third-order probability amplitude coefficient is given by

$$a_L^{(3)}=\frac{h^{-3}H_{LK'}H_{KN'}H_{NM'}}{\nu_{NM}}\times\left[\frac{1}{\nu_{KM}}\left\{\frac{e^{i2\pi\nu_{LM}t}-1}{i\nu_{LM}}-\frac{e^{i2\pi\nu_{LK}t}-1}{i\nu_{LK}}\right\}-\frac{1}{\nu_{KN}}\left\{\frac{e^{i2\pi\nu_{LN}t}-1}{i\nu_{LN}}-\frac{e^{i2\pi\nu_{LK}t}-1}{i\nu_{LK}}\right\}\right].$$

The intermediate states are denoted by the subscripts N and K. We are interested in three photon transitions of the type in which energy is conserved or nearly conserved in all three steps. The subscript N refers to the state in which a particular oscillator excited by the lamp has lost one photon. The subscript K refers to the state in which a particular microwave field oscillator has lost a photon. In order to calculate the probability that the particle has returned to the ground state with emission of a circularly polarized photon we must square the probability amplitude and integrate over all lamp photons, microwave photons, and emitted circularly polarized photons. We denote by $\rho(N)$, $\rho(K)$, and $\rho(L)$ the density of intermediate and final states. The transition probability is then

$$W=\frac{1}{h^6t}\int\left|\frac{H_{LK'}H_{KN'}H_{NM'}}{\nu_{NM}}\right|^2\times\left[\frac{1}{\nu_{KM}}\left\{\frac{e^{i2\pi\nu_{LM}t}-1}{\nu_{LM}}-\frac{e^{i2\pi\nu_{LK}t}-1}{\nu_{LK}}\right\}-\frac{1}{\nu_{KN}}\left\{\frac{e^{i2\pi\nu_{LN}t}-1}{\nu_{KN}}-\frac{e^{i2\pi\nu_{LK}t}-1}{\nu_{LK}}\right\}\right]^2\times\rho(E_L)\rho(E_K)\rho(E_N)dE_LdE_KdE_N.$$

We need to evaluate this expression, under conditions such that energy is conserved or nearly conserved in all steps. All denominators may vanish. An approximate solution can be obtained in the following way. When integrating over dE_N, we are integrating over all lamp photons which can be absorbed. Let the lamp have intensity which is zero except for a range $\Delta\nu_1$ on each side of that required for resonance. Let the lamp intensity be assumed constant over the range $2\Delta\nu_1$. A study of the integrand shows that it is not singular in the range where all denominators vanish. The integral over E_N is along the real frequency axis from $-\Delta\nu_1$ to $+\Delta\nu_1$. We can equally well integrate along a semicircle in the lower half complex frequency plane from $-\Delta\nu_1$ on the real axis to $+\Delta\nu_1$ on the real axis. The second pair of terms which have the factor $1/(\nu_{KN})$ may then be neglected in comparison with the pair of terms which have the factor $1/(\nu_{KM})$. This integration along the semicircle is now readily performed. The integration over E_K can be performed in the same way if the relaxation time for interaction with other particles is much smaller than the spontaneous emission lifetime. This leaves one integration which may then be handled in the same way as for one photon processes. The result

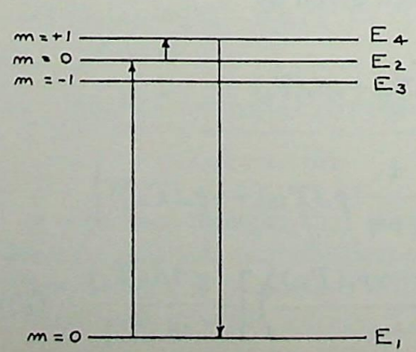


FIG. 19. Energy levels for quantum mechanical amplifier without spontaneous emission noise.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

then gives the transition probability

$$W=\frac{16\pi^2\rho(E_N)\rho(E_K)\rho(E_L)|H_{LK}'H_{KN}'H_{NM}'|^2}{h^3\Delta\nu_{NM}\Delta\nu_{KM}}.$$

All quantities are evaluated at resonance. This predicts modified absorption (or emission) at one frequency due to intense radiation at a different frequency, for single particles. $\rho(N)$ is calculated from the average number of field oscillators excited by the lamp. It is reasonable to suppose that each lamp excited field oscillator is unlikely to have more than one photon. The remaining quantities require, for calculation, a detailed knowledge of the transitions and the experimental arrangements. The excited states of the rare earth salts may be suitable for such counters and are now being investigated by Mr. U. E. Hochuli.

At a meeting held at the National Academy of Sciences in April, 1958, Bloembergen independently proposed a microwave (or infrared) photon counter which operates on similar principles (Fig. 20).

Consider a solid with at least three low-lying levels and one or more optical levels. The solid, such as a salt of the rare earth elements, is cooled to such a low temperature that only the ground state is occupied. $kT\ll E_2-E_1$. Strong optical illumination takes place at the frequency $h^{-1}(E_4-E_2)$. This light is normally not absorbed because the level E_2 is empty, unless it is excited by an incident quantum E_2-E_1. Then the system gets raised to level E_4. The spontaneous emission from level E_4 to E_3 can be detected by a photomultiplier tube. Again, discrimination from the strong incident background can be made by directional, frequency, or polarization filters.

VII. MASERS AT LOW FREQUENCIES AND IN THE INFRARED

There has not been a great deal of motivation for development of masers at low frequencies. The theory previously given would appear to be applicable. Nuclear spin energy levels and the levels associated with nuclear quadrupole spectra are available in this part of the spectrum. The small nuclear moment gives a weaker interaction with radiation, but this tends to be compensated by longer transverse relaxation times. As noted by Braunstein[38] a nucleus of spin $I>2$ which possesses an electric quadrupole moment will yield at least three unequally spaced levels in a crystal of lower than cubic symmetry. If the crystalline field has large deviations from axial symmetry the $\Delta m=2$ transitions (necessary for saturation) begin to have significantly large transition probabilities.

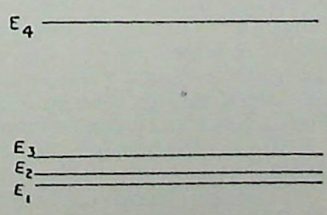


FIG. 20. Energy levels for quantum mechanical amplifier without spontaneous emission noise.

Allais[39] obtained maser action in nuclear resonance experiments in the megacycle region. This followed the work of Abragam, Combrisson, and Solomon[39] in the kilocycle region.

In the infrared and optical parts of the spectrum the sensitivity of a maser will be expected to be less than that of photon counting devices. At 8000 A the spontaneous emission equivalent temperature is 18 000°K. Achievement of coherent amplification would be desirable. This is a useful concept only when large numbers of photons are involved so that the phase can be well defined.[40] The availability of highly monochromatic sources would extend the resolution of spectroscopy in these regions. One problem is to provide a device at infrared and optical frequencies which has the energy storage capacity ordinarily provided by an electromagnetic cavity resonator, with wide mode separations. Dicke has proposed[41] a number of arrangements which employ a standing wave system between parallel planes. Similar proposals have been made by Schawlow and Townes.[42] They suggest that a multimode cavity could be employed. Single modes might be selected by reducing the cavity to partially transparent end plates, with open sides. The great directivity associated with diffraction from sources which are many wavelengths on a side may make it possible to select particular modes and suppress unwanted ones. They carry out calculations for a system using potassium vapor.

VIII. PARAMETRIC AMPLIFIERS

During the past few years another type of microwave amplifier has been developed, which has promise of low noise performance approaching that of a maser. We describe it in classical terms because nothing seems to be gained by a quantum mechanical description. In Fig. 21 two oscillators are coupled by a time-dependent capacitor. Let the natural frequencies of the oscillators be ν_1 and ν_2, and let the capacity C_3 have the time dependence

$$C_3=C_o+C_t\sin[2\pi(\nu_1+\nu_2)t].$$

It has been known since at least the time of Lord Rayleigh that two harmonic oscillators with natural frequencies ν_1 and ν_2, with this kind of time-dependent coupling may be unstable.

[38] R. Braunstein, Phys. Rev. **107**, 1195 (1958).

[39] Abragam, Combrisson, and Solomon, Compt. rend. **245**, 157 (1957); E. Allais, *ibid.* **246**, 2123 (1958).

[40] W. Heitler, *The Quantum Theory of Radiation* (Oxford University Press, London), third edition, p. 65.

[41] R. H. Dicke, U. S. Patent 2,851,652, issued September 9, 1958.

[42] A. L. Schawlow and C. H. Townes, Phys. Rev. **112**, 1940 (1958); see also A. M. Prokhorov, J. Exptl. Theoret. Phys. **34**, 1658 (1958). Research activity is increasing along this line. Since low noise is not essential we suggest review of older methods in conjunction with parallel planes.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

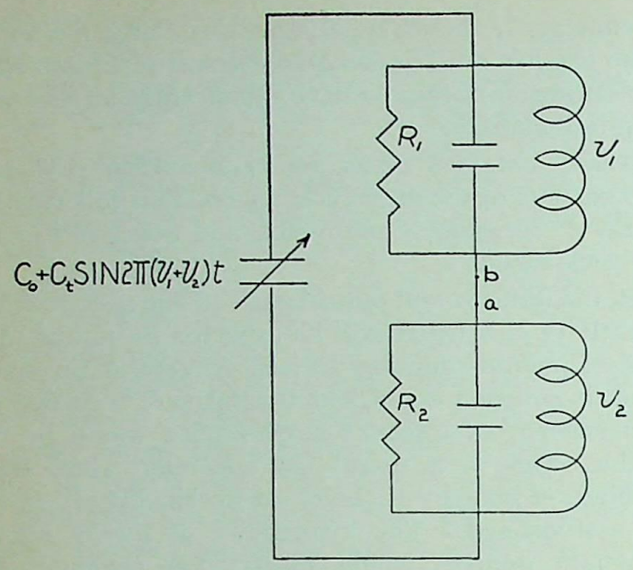


FIG. 21. Parametric amplifier.

Consider first a capacity alone with

$$C=C_t \sin[2\pi(\nu_1+\nu_2)t]$$

driven by a voltage $V=V_1 \sin(2\pi\nu_1 t)+V_2 \sin(2\pi\nu_2 t)$. The principal part[43] of the current is

$$i=-\pi\nu_1 C_t V_2 \sin(2\pi\nu_1 t)-\pi\nu_2 C_t V_1 \sin(2\pi\nu_2 t).$$

If, in the circuit of Fig. 21, a voltage $V_G \sin 2\pi\nu_1 t$ is applied between terminals a and b, the output voltage of frequency ν_1 can be shown to have increased amplitude

$$V_1=V_G\left[1-i2\pi C_o R_1\nu_1-\frac{\pi^2\nu_1\nu_2 C_t^2 R_1 R_2}{1-i2\pi\nu_2 C_o R_2}\right]^{-1}.$$

In deducing this we assume that the oscillator of frequency ν_1 has impedance R_1 at ν_1 and zero impedance at all other frequencies. A corresponding assumption is made concerning the oscillator of frequency ν_2. The time-dependent capacitor or inductance need have very little random fluctuations. The low noise properties of this type of amplifier were pointed out unequivocally by van der Ziel[44] in 1948. This pioneering theoretical work appears to represent the earliest solution to the problem of low noise microwave amplification. Experimental verification of the low noise properties was accomplished by Salzberg and Sard.[45] The noise figure of a parametric amplifier has been given as[43]

$$F=1+\frac{G_1}{G_g}+\frac{\nu_1}{\nu_2}\frac{G_{T1}}{G_g}+F_S+F_G,$$

where G_1 = effective conductance of unloaded amplifying resonant circuit, G_g = effective conductance coupled into the amplifying resonant circuit by the generator, G_{T1} = effective conductance of the loaded amplifying resonant circuit, ν_1 = amplifying frequency, ν_2 = idling frequency, F_S = shot noise, and F_G = gain fluctuation noise.

The revival of interest in this type of amplifier appears to have started with the work of Suhl[46] and Weiss[47] on the ferromagnetic amplifier. Low noise "up frequency conversion" is also possible with parametric devices.[48] Room temperature operation is possible, with moderately good noise performance.

IX. APPLICATIONS OF MASERS TO EXPERIMENTAL PHYSICS AND RADIO ASTRONOMY

The very low noise, coupled with adequate gain-bandwith product, make the low-temperature solid state maser an obvious tool for radio astronomy and radar.‖ A modified Dicke-type radiometer in which a maser is used for pre-amplification has been described.[49] The ruby maser of Alsop, Giordmaine, and Townes, which was described earlier, was mounted near the focus of the Naval Research Laboratory 50-ft reflector in order to minimize transmission line losses. Liquid helium cooling was provided by a stainless steel Dewar with 3 l capacity which was maintained under partial vacuum. This allowed about 15 hr of observation before recharging was needed. The complete radiometer installation made it possible to observe with an rms output fluctuation of 0.03°K for an output time constant of 5 sec. The intermediate frequency amplifiers had a band width of 5 Mc.

The great stability of oscillators such as the cesium beam clock and the molecular beam masers make possible certain experimental tests of both special and general relativity. These possibilities have been discussed by Moller.[50] A molecular beam maser has been employed to repeat the Michelson-Morley experiment.[51] If there is a fixed ether a difference in frequency would

[43] We are following the treatment of H. Heffner and G. Wade, Meeting of Institute of Radio Engineers Professional Group on Electron Devices, Washington, D. C. (October, 1957).

[44] A. van der Ziel, J. Appl. Phys. **19**, 999 (1948).

[45] B. Salzberg and E. W. Sard, Proc. Inst. Radio Engrs. **46**, 1303 (1958).

[46] H. Suhl Phys. Rev. **106**, 384 (1957); J. Appl. Phys. **28**, 1225 (1957).

[47] M. T. Weiss, Phys. Rev. **107**, 317 (1957).

[48] Herrmann, Uenohara, and Uhlir, Proc. Inst. Radio Engrs. **46**, 1301(L) (1958).

‖ *Note added in proof.*—Operation of a ruby maser at 60°K was reported by C. R. Ditchfield and P. A. Forrester [Phys. Rev. Letters **1**, 448 (1958)], and independently by T. H. Maiman (Hughes Aircraft Company Report). A maser employing a slow wave structure with band width of 67 mc per sec and gain of 13 decibels was described by R. W. DeGrasse (Bell Telephone Laboratories Report entitled "Slow Wave Structures for Unilateral Solid State Maser Amplifiers"). Amplifiers employing ruby which have wide tunability were reported by G. K. Wessel [380–450 mc, Proc. Inst. Radio Engrs. **47**, 590 (1959)], and by F. R. Arams and S. Okwit [850–2000 mc, Proc. Inst. Radio Engrs. **47**, 992 (1959)]. A proposal for generating higher frequencies than the pump frequency was made by A. E. Siegman and R. J. Morris [Phys. Rev. Letters **2**, 302 (1959)]. They suggest a staircase scheme in which say the spin populations of the lowest pair of states are reversed by adiabatic rapid passage, followed by adiabatic rapid passage reversal of the populations of levels 2 and 3. Then levels 3 and 1 can be employed for power generation.

[49] Mayer, McCullough, and Sloanmaker, Astrophys. J. **127**, 1 (1958).

[50] C. Moller, Nuovo cimento **6**, Suppl., 381 (1957).

[51] Cedarholm, Bland, Havens, and Townes, Phys. Rev. Letters **1**, 342 (1958).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

be expected between two molecular beam type masers with beam velocities parallel and antiparallel to the earth's orbital motion. Let u be the velocity of the molecules relative to the cavity which is assumed to be moving with velocity v with respect to the fixed ether (u is parallel to v). It is assumed that the photons radiated by the molecules move with a velocity **c** relative to the fixed ether. The photon velocity relative to the cavity must be normal to u. This requires that **c** be tilted forward of the normal to u by an angle $\phi \approx (v/c)$. The expected doppler shift would be $vu\phi/c = uv v/c^2$. For a thermal velocity of 0.6 kMs/sec and for the earth's orbital velocity the difference in frequency between two *oppositely* directed beams is $2uvv/c^2$, which is ~ 10 cps if $\nu = 23\,870$ Mc/sec. The experiment was done by mounting two maser oscillators on a rack which could be rotated about a vertical axis. The oscillators were adjusted so that their frequencies differed by a small amount. This difference was recorded and the apparatus was then rotated through 180°. A slight change of about 1/50 cps was observed under the best conditions of operation. This is smaller by a factor of 1000 than what would be expected on the basis of a fixed ether. Inasmuch as the special theory of relativity is one of the most securely established of all physical theories, such experiments may be regarded as a search for other effects such as perhaps an anisotropy of space in this part of our galaxy. The use of stable oscillators to test the gravitational red shift has also been considered.[50,52] This is perhaps more a test of whether or not these devices are natural clocks (whose intervals are invariants) than a test of the fundamental postulates of the general theory of relativity.

A series of experiments is being planned by the author to search for gravitational radiation. The theory of an antenna for such radiation has been discussed.[53] Two masses which are separated will have forces exerted upon them by a gravitational wave. The phase difference at the two masses results in one being driven relative to the other. Also strains can be set up in a material by a gravitational wave. Under some conditions it is desirable to make use of acoustical resonance, under other conditions where acoustic phase reversal is troublesome, it is better not to employ acoustical resonance. The gravitational wave interacts both with the mass of the piezoelectric crystal and the conducting masses shown. Very low-frequency search is planned. The output voltages are amplified as shown in Fig. 22. Liquid helium temperatures and low noise receivers may make it possible to observe correlation in the outputs, in the presence of electrical noise.

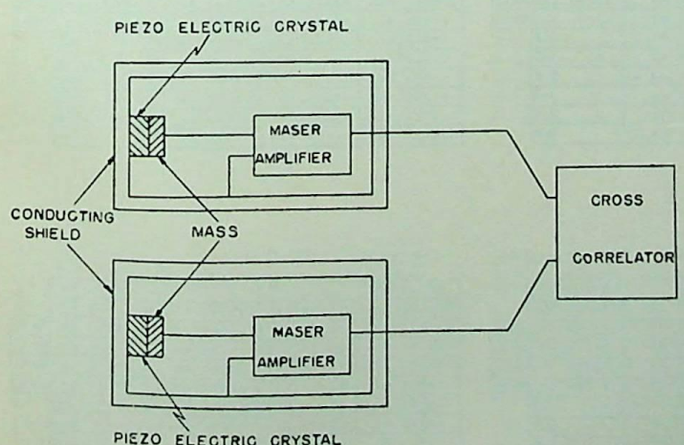


FIG. 22. Masers for detection of gravitational radiation at radio-frequencies.

X. CONCLUSION

The new microwave amplifiers are the result of electronics research to develop millimeter wave techniques, magnetic resonance research, and research in microwave spectroscopy. Maser amplifiers bring to the microwave region a detection sensitivity of the order of a few microwave photons. Most of these amplifiers employ the magnetic moment of the electron, and therefore fill in the gap in modern electronics pointed out by Sommerfeld. We may look forward to important advances in radio astronomy, spectroscopy and solid state physics, in consequence of improved ability to distinguish weak signals.

ACKNOWLEDGMENTS

We thank Professor Bloembergen, Professor Strandberg, and Professor Townes for helpful criticism and a profitable correspondence. Suggestions for improving the manuscript were also made by Dr. L. S. Nergaard and Dr. E. O. Schulz-DuBois. We acknowledge stimulating discussions with Dr. R. K. Wangsness and Professor R. A. Ferrell.

APPENDIX. MACHINE CALCULATIONS OF MATRIX ELEMENTS AND ENERGY LEVELS FOR RUBY¶

The following data are entirely the work of W. S. Chang and A. E. Siegman of the Stanford University Electronics Laboratories. We thank them for their kindness and cooperation in allowing this work to be published here.

The spin Hamiltonian used was

$$\mathcal{H} = g\beta_M \mathbf{H}_{dc} \cdot \mathbf{S} + D(S_z^2 - 5/4),$$

with $g = 1.99$ and $2D = +11.46$ kmc/s. The small anisotropy in the g tensor was ignored.

[52] S. F. Singer, Phys. Rev. **104**, 11 (1956). J. R. Zacharias has considered terrestrial tests employing a cesium beam clock.

[53] J. Weber, Gravity Research Foundation Essay, New Boston, New Hampshire (April, 1958) and April, 1959.

¶ This material originally appeared as Tech. Rept. No. 156–2 under Air Force Contract AF33(600)-27784 of the Stanford Electronics Laboratories, Stanford University, California. A similar set of calculations have also been performed for the material potassium chromicyanide, $K_3Cr(CN)_6$, and reported earlier in Tech. Rept. No. 156–1 under the same contract. Because of the lower symmetry of the chromicyanide, the results are too lengthy to present here. Copies of the report can be obtained directly from the Stanford Electronics Laboratories. The support of this work by the Wright Air Development Center is gratefully acknowledged.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

TABLE I(a). Ruby $\theta = 10°$.

17453 49			50000 49	
64626 50N	50051 50N	36786 50	77941 50	
24804 47N	16853 49	98548 49 •	20405 48	for E_1
16158 48	98556 49N	16830 49 •	40671 47	for E_2
29417 45N	57731 46	20799 48	99978 49	for E_3
99986 49N	16344 48N	25300 46 •	17341 47	for E_4
18817 50	20007 50N	33252 49 •	1-2	
33277 49N	33277 49N	92644 47 •	2-3	
20548 47N	21808 47	51172 47 •	3-4	
16999 50N	16991 50N	40608 48 •	1-3	
17118 50	17127 50N	32422 48 •	2-4	
26263 49N	25668 49	43304 47	1-4	

17453 49		10000 50		
72064 50N	42862 50N	16288 50	98638 50	
41318 47N	16798 49	98466 49 •	46866 48	for E_1
89242 48	98534 49N	16764 49 •	11935 48	for E_2
19840 46N	38882 47	48205 48	99882 49	for E_3
99956 49N	29520 48N	82463 46 •	73835 46	for E_4
18657 50	20049 50N	33263 49 •	1-2	
38480 49N	38487 49N	31531 48 •	2-3	
10939 48N	11069 48	16925 47 •	3-4	
16911 50N	16835 50N	92312 48 •	1-3	
17157 50	17156 50N	58765 48 •	2-4	
23407 49N	23142 49	65152 47	1-4	

17453 49		15000 50		
79668 50N	35733 50N	39737 49 •	11937 51	
52844 47N	16691 49	98251 49 •	82286 48	for E_1
40030 48	98496 49N	16476 49 •	33220 48	for E_2
97454 46N	19060 48	86662 48	99603 49	for E_3
99918 49	40362 48	14941 47	23014 46	for E_4
18210 50	20260 50N	33512 49 •	1-2	
45328 49N	45401 49N	10387 49 •	2-3	
42584 48	42537 48N	29690 47 •	3-4	
16916 50N	16409 50N	15754 49 •	1-3	
17180 50N	17176 50	80461 48	2-4	
20950 49	21003 49N	76920 47 •	1-4	

17453 49			20000 50	
87539 50N	28974 50N	23629 50 •	14014 51	
60918 47N	16498 49	97748 49 •	13142 49	for E_1
47405 48	95305 49N	12460 49 •	27186 49	for E_2
12793 48N	24904 49	17029 49	95331 49	for E_3
99877 49	49432 48	21514 47	58699 46	for E_4
13570 50	23397 50N	38710 49 •	1-2	
47065 49N	54913 49N	99548 49 •	2-3	
45137 49	45107 49N	28066 48 •	3-4	
20122 50N	11086 50N	16809 49 •	1-3	
16601 50N	16598 50	94719 48	2-4	
18046 49	19007 49N	82257 47 •	1-4	

17453 49		25000 50		
93864 50N	43740 50N	21331 50 •	16093 51	
66087 47N	16141 49	96594 49 •	20212 49	for E_1
57094 47	11108 49N	18568 49	97629 49	for E_2
56523 48N	97894 49	18018 49	77442 48	for E_3
99836 49	57128 48	27512 47	28468 46	for E_4
12909 50N	18533 50N	43067 49 •	1-2	
67367 49	60455 49N	36998 49 •	2-3	
17136 50	17138 50N	11392 49 •	3-4	
19265 50N	17613 50	28622 49	1-3	
16617 49N	16626 49	94096 47	2-4	
16947 49	16996 49N	80871 47 •	1-4	

17453 49			30000 50	
10500 51N	62448 50N	14294 50 •	18174 51	
68019 47N	15452 49	93942 49 •	30586 49	for E_1
48383 47	93557 48N	29424 49	95112 49	for E_2
63276 48N	98148 49	17573 49	42498 48	for E_3
99796 49	63739 48	33155 47	22405 46	for E_4
11226 50N	17885 50N	61087 49 •	1-2	
87435 49	88034 49N	26571 49 •	2-3	
17167 50	17195 50N	12749 49 •	3-4	
17943 50N	17414 50	27904 49	1-3	
11871 49N	11877 49	69073 47	2-4	
14900 49	14918 49N	76059 47 •	1-4	

17453 49			35000 50	
11562 51N	79711 50N	72369 49 •	20257 51	
65428 47N	14131 49	88111 49 •	45125 49	for E_1
55808 47	10689 49N	43954 49	89181 49	for E_2
69049 48N	98172 49	17441 49	32132 48	for E_3
99757 49	69477 48	38421 47	34837 46	for E_4
70231 49N	17608 50N	83522 49 •	1-2	
11419 50	11485 50N	26872 49 •	2-3	
17200 50	17213 50N	13914 49 •	3-4	
16214 50N	15997 50	25027 49	1-3	
11756 49N	11779 49	66547 47	2-4	
12450 49	12497 49N	68489 47 •	1-4	

17453 49			40000 50	
12861 51N	94608 50N	18315 48 •	22341 51	
57376 47N	11992 49	77641 49 •	61867 49	for E_1
68069 47	12847 49N	60576 49	78517 49	for E_2
74094 48N	98161 49	17381 49	27167 48	for E_3
99721 49	74502 48	43117 47	20662 46	for E_4
59294 48N	17593 50N	10025 50 •	1-2	
14375 50	14466 50N	29690 49 •	2-3	
17205 50	17225 50N	14929 49 •	3-4	
13580 50N	13488 50	20351 49	1-3	
12579 49N	12576 49	76933 47	2-4	
96076 48	96045 48N	52662 47 •	1-4	

17453 49			45000 50	
14435 51N	10676 51N	68636 49	24425 51	
46566 47N	95477 48	64794 49 •	75566 49	for E_1
79205 47	14663 49N	74165 49	65451 49	for E_2
78548 48N	98140 49	17346 49	24255 48	for E_3
99686 49	78940 48	47503 47	27320 46	for E_4
57030 49	17537 50N	10173 50 •	1-2	
16660 50	16785 50N	32204 49 •	2-3	
17207 50	17234 50N	15827 49 •	3-4	
10597 50N	10557 50	15220 49	1-3	
13065 49N	13073 49	80524 47	2-4	
68736 48	68808 48N	39305 47 •	1-4	

17453 49			50000 50	
16218 51N	11683 51N	13903 50	26511 51	
37184 47N	75430 48	53728 49 •	84000 49	for E_1
86825 47	15735 49N	82539 49	54211 49	for E_2
82509 48N	98116 49	17324 49	22341 48	for E_3
99654 49	82887 48	51465 47	23836 46	for E_4
99216 49	17500 50N	93462 49 •	1-2	
17970 50	18126 50N	33587 49 •	2-3	
17207 50	17241 50N	16625 49 •	3-4	
81402 49N	81207 49	11171 49	1-3	
12997 49N	12991 49	82791 47	2-4	
48638 48	48596 48N	27000 47 •	1-4	

17453 49			55000 50	
18125 51N	12565 51N	20937 50	28597 51	
30440 47N	61370 48	45688 49 •	88739 49	for E_1
91833 47	16301 49N	87250 49	46052 49	for E_2
86054 48N	98091 49	17309 49	20989 48	for E_3
99624 49	86420 48	55099 47	21835 46	for E_4
12345 50	17474 50N	83745 49 •	1-2	
18649 50	18831 50N	34230 49 •	2-3	
17207 50	17247 50N	17339 49 •	3-4	
63995 49N	63808 49	84195 48	1-3	
12618 49N	12600 49	82495 47	2-4	
35445 48	35350 48N	18580 47 •	1-4	

17453 49			60000 50	
20100 51N	13301 51N	27965 50	30405 51	
25772 47N	51759 48	40041 49 •	91486 49	for E_1
95370 47	16608 49N	89983 49	40326 49	for E_2
89247 48N	98066 49	17299 49	19981 48	for E_3
99596 49	89601 48	58497 47	30924 46	for E_4
13739 50	17455 50N	75515 49 •	1-2	
19012 50	19215 50N	34526 49 •	2-3	
17206 50	17251 50N	17982 49 •	3-4	
51955 49N	51890 49	65877 48	1-3	
12114 49N	12118 49	79763 47	2-4	
26793 48	26810 48N	15690 47 •	1-4	

17453 49		65000 50		
22112 51N	14158 51N	34988 50	32772 51	
22478 47N	45030 48	36000 49 •	93185 49	for E_1
98073 47	16785 49N	91676 49	36231 49	for E_2
92136 48N	98043 49	17291 49	19201 48	for E_3
99569 49	92479 48	61619 47	38302 46	for E_4
14586 50	17440 50N	69045 49 •	1-2	
19220 50	19438 50N	34667 49 •	2-3	
17204 50	17255 50N	18363 49 •	3-4	
43424 49N	43380 49	53340 48	1-3	
11594 49N	11616 49	77066 47	2-4	
20974 48	21061 48N	14027 47 •	1-4	

17453 49		70000 50		
24148 51N	14913 51N	42007 50	34060 51	
20079 47N	40150 48	33016 49 •	94306 49	for E_1
10026 48	16096 49N	92794 49	33208 49	for E_2
94763 48N	98020 49	17285 49	18582 48	for E_3
99544 49	95095 48	64457 47	38247 46	for E_4
15131 50	17428 50N	63999 49 •	1-2	
19347 50	19578 50N	34737 49 •	2-3	
17202 50	17259 50N	19091 49 •	3-4	
37165 49N	37134 49	46421 48	1-3	
11103 49N	11121 49	75124 47	2-4	
16927 48	16991 48N	11321 47 •	1-4	

17453 49			75000 50	
26198 51N	15653 51N	49022 50	36948 51	
18272 47N	36487 48	30742 49 •	95087 49	for E_1
10210 48	16968 49N	93572 49	30903 49	for E_2
97163 48N	97998 49	17280 49	10075 48	for E_3
99521 49	97484 48	67097 47	40659 46	for E_4
15503 50	17418 50N	60020 49 •	1-2	
32419 49N	32596 49	37851 48	2-3	
13974 48	14044 48N	99553 46 •	3-4	
19430 50	19670 50N	34772 49 •	1-3	
10636 49N	10657 49	72886 47	2-4	
17200 50	17262 50N	19574 49 •	1-4	

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

TABLE I(b). Ruby $\theta = 20°$.

34906 49			50000 50	
65568 50N	49333 50N	37805 50	77096 50	
89413 47N	30842 49	95044 49 •	38052 48	for E_1
31103 48	95069 49N	30821 49 •	14643 48	for E_2
23034 46N	21029 47	40713 48	99916 49	for E_3
99947 49	32345 48	11070 47	41471 46	for E_4
16039 50	20029 50N	58867 49 •	1-2	
61063 49N	61076 49N	33442 48 •	2-3	
83498 47	83075 47N	19082 47 •	3-4	
16213 50N	16191 50N	74688 48 •	1-3	
16662 50N	16651 50	62833 48	2-4	
47202 49	47302 49N	15828 48 •	1-4	

34906 49			10000 50	
74222 50N	41641 50N	18752 50	97111 50	
14771 48N	30535 49	96831 49 •	85000 48	for E_1
56798 48	95031 49N	30315 49 •	42024 48	for E_2
14878 47N	14020 48	93768 48	99549 49	for E_3
99827 49	58609 48	33527 47	28799 46	for E_4
15579 50	20176 50N	59093 49 •	1-2	
69877 49N	70032 49N	11065 49 •	2-3	
41003 48	41024 48N	48084 47 •	3-4	
15917 50N	15646 50N	16058 49 •	1-3	
16792 50N	16776 50	11538 49	2-4	
41744 49	41729 49N	23233 48 •	1-4	

34906 49			15000 50	
83397 50N	34529 50N	43895 48	11728 51	
18622 48N	29999 49	94296 49 •	14305 49	for E_1
78067 48	94857 49N	28679 49 •	10891 49	for E_2
65644 47N	61350 48	16088 49	98370 49	for E_3
99675 49	80302 48	59583 47	64534 46	for E_4
14390 50	20778 50N	60611 49 •	1-2	
79330 49N	80821 49N	33272 49 •	2-3	
14319 49	14318 49N	17613 48 •	3-4	
15845 50N	14300 50N	24415 49 •	1-3	
16833 50N	16818 50	15076 49	2-4	
36788 49	36802 49N	26254 48 •	1-4	

34906 49			20000 50	
93284 50N	28764 50N	15536 50 •	13758 51	
21020 48N	29155 49	93203 49 •	21415 49	for E_1
89064 48	88761 49N	19492 49 •	41232 49	for E_2
37388 48N	34269 49	31150 49	88550 49	for E_3
99510 49	98478 48	85679 47	90722 46	for E_4
88982 49	23721 50N	70165 49 •	1-2	
63055 49N	93695 49N	13519 50 •	2-3	
66513 49	66516 49N	82947 48 •	3-4	
18024 50N	71820 49N	17630 49 •	1-3	
15604 50N	15598 50	17893 49	2-4	
32205 49	32216 49N	26638 48 •	1-4	

34906 49			25000 50	
10413 51N	37067 50N	16782 50 •	15798 51	
22204 48N	27917 49	91235 49 •	29963 49	for E_1
42619 48	39749 49N	17202 49	90032 49	for E_2
10385 49N	86565 49	37134 49	31658 49	for E_3
99343 49	11389 49	11052 48	11805 47	for E_4
47650 49N	21462 50N	76345 49 •	1-2	
81185 49	10856 50N	12767 50 •	2-3	
15813 50	15824 50N	21604 49 •	3-4	
17658 50N	11132 50	30403 49	1-3	
62755 49N	62750 49	76655 48	2-4	
27817 49	27832 49N	25353 48 •	1-4	

34906 49			30000 50	
11621 51N	51996 50N	10241 50 •	17845 51	
22310 48N	26232 49	88093 49 •	39323 49	for E_1
31530 48	26905 49N	31622 49	90302 49	for E_2
12185 49N	91185 49	35181 49	17293 49	for E_3
99179 49	12712 49	13352 48	13702 47	for E_4
46180 49N	19397 50N	96467 49 •	1-2	
11291 50	12466 50N	85483 49 •	2-3	
16582 50	16615 50N	25205 49 •	3-4	
15368 50N	13077 50	35327 49	1-3	
39528 49N	39523 49	49135 48	2-4	
23575 49	23575 49N	22889 48 •	1-4	

34906 49			35000 50	
12974 51N	66101 50N	31387 49 •	19898 51	
21519 48N	24156 49	83723 49 •	49012 49	for E_1
30510 48	27536 49N	42441 49	86204 49	for E_2
13435 49N	92011 49	34449 49	12906 49	for E_3
99022 49	13657 49	15470 48	16114 47	for E_4
20062 49N	18689 50N	98067 49 •	1-2	
13132 50	14056 50N	74563 49 •	2-3	
16742 50	16798 50N	27757 49 •	3-4	
13594 50N	12494 50	32959 49	1-3	
33156 49N	33153 49	41743 48	2-4	
19538 49	19543 49N	19867 48 •	1-4	

34906 49			40000 50	
14478 51N	76841 50N	40509 49	21957 51	
20130 48N	21886 49	78484 49 •	57939 49	for E_1
31652 48	28099 49N	51718 49	80781 49	for E_2
14483 49N	92253 49	34092 49	10830 49	for E_3
98874 49	14858 49	17406 48	18325 47	for E_4
11013 49	18340 50N	10616 50 •	1-2	
14577 50	15457 50N	71176 49 •	2-3	
16803 50	16885 50N	29906 49 •	3-4	
11910 50N	11307 50	28997 49	1-3	
30316 49N	30317 49	38695 48	2-4	
15899 49	15910 49N	16716 48 •	1-4	

34906 49			45000 50	
16117 51N	90270 50N	11251 50	24019 51	
18500 48N	19672 49	73012 49 •	65412 49	for E_1
33273 48	29038 49N	59305 49	75023 49	for E_2
15393 49N	92314 49	33867 49	96257 48	for E_3
98734 49	15739 49	19165 48	19750 47	for E_4
40117 49	18133 50N	10982 50 •	1-2	
15680 50	16574 50N	70105 49 •	2-3	
16831 50	16940 50N	31772 49 •	3-4	
10345 50N	99856 49	24867 49	1-3	
28502 49N	28496 49	37041 48	2-4	
12830 49	12830 49N	13717 48 •	1-4	

34906 49			50000 50	
17067 51N	10061 51N	18442 50	26085 51	
16901 48N	17695 49	67861 49 •	71265 49	for E_1
34843 48	29897 49N	65199 49	69591 49	for E_2
16195 49N	92304 49	33756 49	88405 48	for E_3
98603 49	16521 49	20773 48	21343 47	for E_4
64277 49	17997 50N	11000 50 •	1-2	
16480 50	17406 50N	69755 49 •	2-3	
16844 50	16978 50N	33417 49 •	3-4	
89721 49N	87428 49	21161 49	1-3	
27018 49N	27010 49	35675 48	2-4	
10363 49	10362 49N	11227 48 •	1-4	

34906 49			55000 50	
19701 51N	11014 51N	25618 50	28154 51	
15474 48N	16022 49	63324 49 •	75702 49	for E_1
36213 48	30575 49N	69653 49	64810 49	for E_2
16910 49N	92260 49	33666 49	82886 48	for E_3
96481 49	17218 49	22245 48	23033 47	for E_4
83113 49	17900 50N	10814 50 •	1-2	
17047 50	18007 50N	69611 49 •	2-3	
16848 50	17008 50N	34881 49 •	3-4	
78171 49N	76620 49	18062 49	1-3	
25673 49N	25668 49	34375 48	2-4	
84392 48	84418 48N	92751 47 •	1-4	

34906 49			60000 50	
21595 51N	11907 51N	32779 50	30225 51	
14259 48N	14644 49	59464 49 •	79041 49	for E_1
37380 48	31084 49N	73002 49	60748 49	for E_2
17550 49N	92200 49	33600 49	78797 48	for E_3
98366 49	17844 49	23591 48	24486 47	for E_4
97441 49	17827 50N	10534 50 •	1-2	
17446 50	18437 50N	69518 49 •	2-3	
16846 50	17033 50N	36190 49 •	3-4	
68654 49N	67551 49	15548 49	1-3	
24427 49N	24423 49	33155 48	2-4	
69563 48	69602 48N	77257 47 •	1-4	

34906 49			65000 50	
23534 51N	12756 51N	39925 50	32298 51	
13244 48N	13518 49	56227 49 •	81571 49	for E_1
38368 48	31463 49N	75341 49	37347 49	for E_2
18128 49N	92132 49	33551 49	75647 48	for E_3
98259 49	18408 49	24825 48	25673 47	for E_4
10830 50	17771 50N	10226 50 •	1-2	
17731 50	18749 50N	69435 49 •	2-3	
16842 50	17053 50N	37369 49 •	3-4	
60857 49N	60038 49	13528 49	1-3	
23272 49N	23266 49	32011 48	2-4	
58111 48	58134 48N	64793 47 •	1-4	

34906 49			70000 50	
25505 51N	13573 51N	47056 50	34373 51	
12399 48N	12595 49	53520 49 •	83518 49	for E_1
39214 48	31747 49N	77496 49	54307 49	for E_2
18651 49N	92062 49	33513 49	73145 48	for E_3
98159 49	18920 49	25960 48	26774 47	for E_4
11661 50	17726 50N	99239 49 •	1-2	
17937 50	18979 50N	69354 49 •	2-3	
16835 50	17070 50N	38435 49 •	3-4	
54448 49N	53818 49	11899 49	1-3	
22202 49N	22194 49	30912 48	2-4	
49172 48	49183 48N	54957 47 •	1-4	

34906 49			75000 50	
27499 51N	14367 51N	54176 50	36449 51	
11695 48N	11835 49	51247 49 •	85042 49	for E_1
39946 48	31962 49N	79027 49	52185 49	for E_2
19180 49N	91991 49	33482 49	71113 48	for E_3
98065 49	19386 49	27006 48	27751 47	for E_4
12307 50	17689 50N	96411 49 •	1-2	
18091 50	19152 50N	69376 49 •	2-3	
16827 50	17083 50N	39404 49 •	3-4	
49138 49N	48638 49	10574 49	1-3	
21215 49N	21204 49	29872 48	2-4	
42109 48	42104 48N	47006 47 •	1-4	

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

TABLE I(c). Ruby $\theta = 30°$.

52359 49			50000 49	
66856 50N	48388 50N	39534 50	75709 50	
17546 48N	41329 49	90895 49 •	61782 48	for E_1
44365 48	90933 49N	41270 49 •	28467 48	for E_2
75341 46N	44760 47	58917 48	99826 49	for E_3
99886 49	47653 48	24248 47	79319 46	for E_4
12895 50	20061 50N	75770 49 •		1-2
81998 49N	82077 49N	65108 48 •		2-3
17592 48	17511 48N	45626 47 •		3-4
15264 50N	15196 50N	99631 48 •		1-3
16137 50N	16111 50	90544 48		2-4
62792 49	62930 49N	30804 48 •		1-4

52359 49			10000 50	
77133 50N	40154 50N	22691 50	94596 50	
28808 48N	40699 49	90614 49 •	11148 49	for E_1
82129 48	90890 49N	40110 49 •	79164 48	for E_2
45562 47N	26721 48	13402 49	99060 49	for E_3
99619 49	86824 48	74774 47	12364 47	for E_4
12222 50	20330 50N	76694 49 •		1-2
92392 49N	93215 49N	20691 49 •		2-3
82211 48	82208 48N	15973 48 •		3-4
14727 50N	14212 50N	19860 49 •		1-3
16390 50N	16348 50	16924 49		2-4
54615 49	54652 49N	44396 48 •		1-4

52359 49			15000 50	
88253 50N	32964 50N	73738 49	11364 51	
35972 48N	39685 49	89950 49 •	17877 49	for E_1
11334 49	90461 49N	36680 49 •	18516 49	for E_2
17314 48N	99647 48	23660 49	96630 49	for E_3
99275 49	11940 49	13415 48	22440 47	for E_4
10686 50	21151 50N	80050 49 •		1-2
99263 49N	10525 50N	54597 49 •		2-3
25032 49	25030 49N	49350 48 •		3-4
14418 50N	12060 50N	26391 49 •		1-3
16410 50N	16365 50	23323 49		2-4
47143 49	47200 49N	48886 48 •		1-4

52359 49			20000 50	
10035 51N	28707 50N	43124 49 •	13337 51	
40209 48N	38298 49	88783 49 •	25188 49	for E_1
12987 49	85189 49N	24775 49 •	44274 49	for E_2
58677 48N	32565 49	38727 49	86053 49	for E_3
98897 49	14676 49	19414 48	31536 47	for E_4
67946 49	23020 50N	89646 49 •		1-2
73366 49N	11780 50N	13473 50 •		2-3
70029 49	70060 49N	14195 49 •		3-4
14975 50N	60112 49N	17463 49 •		1-3
15185 50N	15156 50	26091 49		2-4
40370 49	40410 49N	48236 48 •		1-4

52359 49			25000 50	
11353 51N	32817 50N	67653 49 •	15311 51	
42261 48N	36582 49	87010 49 •	32756 49	for E_1
10690 49	63971 49N	12347 48	76103 49	for E_2
12775 49N	65425 49	49206 49	55991 49	for E_3
98512 49	16996 49	25146 48	40838 47	for E_4
96914 48	22790 50N	98469 49 •		1-2
30988 49	13027 50N	17439 50 •		2-3
13067 50	13092 50N	28559 49 •		3-4
14720 50N	40860 49	13346 49		1-3
10548 50N	10535 50	19759 49		2-4
34249 49	34289 49N	44852 48 •		1-4

52359 49			30000 50	
12784 51N	43240 50N	19503 49 •	17303 51	
42710 48N	34631 49	84674 49 •	40158 49	for E_1
86546 48	49972 49N	18994 49	84065 49	for E_2
16633 49N	77091 49	49600 49	36332 49	for E_3
90133 49	18981 49	30467 48	49224 47	for E_4
34071 48N	20946 50N	10361 50 •		1-2
89839 49	14196 50N	14389 50 •		2-3
15143 50	15212 50N	36383 49 •		3-4
13066 50N	81397 49	27057 49		1-3
73330 49N	73255 49	14173 49		2-4
28823 49	28855 49N	40137 48 •		1-4

52359 49			35000 50	
14325 51N	54571 50N	47278 49	19310 51	
42074 48N	32579 49	81926 49 •	46999 49	for E_1
80022 48	45178 49N	29670 49	83752 49	for E_2
18954 49N	80431 49	48941 49	27859 49	for E_3
97769 49	20693 49	35343 48	57004 47	for E_4
64746 48	19891 50N	10951 50 •		1-2
11480 50	15231 50N	12540 50 •		2-3
15754 50	15879 50N	41158 49 •		3-4
11615 50N	89139 49	29409 49		1-3
59131 49N	59066 49	11641 49		2-4
24126 49	24154 49N	35074 48 •		1-4

52359 49			40000 50	
15965 51N	63491 50N	11870 50	21327 51	
40799 48N	30557 49	78981 49 •	53023 49	for E_1
78777 48	43650 49N	37388 49	81453 49	for E_2
20726 49N	81662 49	48458 49	23524 49	for E_3
97425 49	22182 49	39702 48	63919 47	for E_4
21963 49	19295 50N	11434 50 •		1-2
12917 50	16102 50N	11614 50 •		2-3
15999 50	16185 50N	44841 49 •		3-4
10380 50N	87126 49	28343 49		1-3
51491 49N	51426 49	10292 49		2-4
20167 49	20188 49N	30212 48 •		1-4

52359 49			45000 50	
17689 51N	75798 50N	19151 50	23354 51	
39226 48N	28666 49	76042 49 •	58141 49	for E_1
79475 48	43300 49N	43376 49	78615 49	for E_2
22197 49N	82168 49	48134 49	20943 49	for E_3
97101 49	23486 49	43014 48	70251 47	for E_4
37747 49	18921 50N	11748 50 •		1-2
13883 50	16811 50N	11114 50 •		2-3
16115 50	16363 50N	47914 49 •		3-4
93084 49N	81943 49	26238 49		1-3
46517 49N	46450 49	94247 48		2-4
16896 49	16914 49N	25878 48 •		1-4

52359 49			50000 50	
19483 51N	85529 50N	26469 50	25388 51	
37579 48N	26959 49	73251 49 •	62396 49	for E_1
80876 48	43386 49N	48128 49	75735 49	for E_2
23462 49N	82364 49	47909 49	19243 49	for E_3
96798 49	24635 49	47476 48	75870 47	for E_4
52038 49	18666 50N	11912 50 •		1-2
14576 50	17375 50N	10820 50 •		2-3
16171 50	16484 50N	50556 49 •		3-4
83789 49N	75900 49	23922 49		1-3
42836 49N	42763 49	87888 48		2-4
14233 49	14245 49N	22137 48 •		1-4

52359 49			55000 50	
21331 51N	94772 50N	33787 50	27430 51	
35989 48N	25454 49	70688 49 •	65895 49	for E_1
82480 48	43621 49N	51938 49	73017 49	for E_2
24570 49N	82407 49	47746 49	18043 49	for E_3
96515 49	25656 49	50807 48	81083 47	for E_4
64349 49	18482 50N	11962 50 •		1-2
15090 50	17820 50N	10635 50 •		2-3
16197 50	16572 50N	52866 49 •		3-4
75764 49N	69921 49	21710 49		1-3
39882 49N	39807 49	82754 48		2-4
12074 49	12085 49N	19015 48 •		1-4

52359 49			60000 50	
23223 51N	10362 51N	41091 50	29476 51	
34518 48N	24143 49	68301 49 •	68769 49	for E_1
84080 48	43890 49N	55018 49	70539 49	for E_2
25552 49N	82370 49	47623 49	17152 49	for E_3
96252 49	26567 49	53841 48	85800 47	for E_4
74719 49	18343 50N	11937 50 •		1-2
15478 50	18171 50N	10509 50 •		2-3
16204 50	16640 50N	54906 49 •		3-4
60857 49N	64368 49	19711 49		1-3
37398 49N	37319 49	70404 48		2-4
10325 49	10335 49N	16416 48 •		1-4

52359 49			65000 50	
25149 51N	11215 51N	48376 50	31527 51	
33188 48N	23007 49	66327 49 •	71135 49	for E_1
85594 48	44149 49N	57530 49	68321 49	for E_2
26428 49N	82289 49	47526 49	16466 49	for E_3
96006 49	27384 49	56612 48	90092 47	for E_4
83383 49	18233 50N	11865 50 •		1-2
15777 50	18449 50N	10420 50 •		2-3
16200 50	16695 50N	56721 49 •		3-4
62912 49N	59358 49	17948 49		1-3
35248 49N	35165 49	74594 48		2-4
89038 48	89121 48N	4261 48 •		1-4

52359 49			70000 50	
27103 51N	12043 51N	55640 50	33503 51	
32000 48N	22024 49	64509 49 •	73097 49	for E_1
86994 48	44382 49N	59599 49	66350 49	for E_2
27216 49N	82184 49	47449 49	15920 49	for E_3
95777 49	28121 49	59148 48	93993 47	for E_4
90616 49	18146 50N	11766 50 •		1-2
16011 50	18671 50N	10354 50 •		2-3
16189 50	16741 50N	58346 49 •		3-4
57782 49N	54897 49	16409 49		1-3
33351 49N	33266 49	71193 48		2-4
77405 48	77474 48N	12468 48 •		1-4

52359 49			75000 50	
29078 51N	12852 51N	62803 50	35641 51	
30944 48N	21171 49	62900 49 •	74737 49	for E_1
88277 48	44587 49N	61320 49	64404 49	for E_2
27929 49N	82066 49	47306 49	15477 49	for E_3
95563 49	28788 49	61476 48	97359 47	for E_4
96677 49	18073 50N	11634 50 •		1-2
16197 50	18851 50N	10304 50 •		2-3
16173 50	16780 50N	59810 49 •		3-4
53338 49N	50945 49	15071 49		1-3
31659 49N	31571 49	68114 48		2-4
67813 48	67870 48N	10971 48 •		1-4

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

TABLE I(d). Ruby $\theta = 40°$.

69813 49			50000 49	
60259 50N	47405 50N	41852 50	73811 50	
26912 48N	49058 49	86874 49 •	61592 48	for E_1
55900 48	86910 49N	48958 49 •	42875 48	for E_2
17264 47N	69979 47	74705 48	99717 49	for E_3
99807 49	51920 48	41262 47	10057 47	for E_4
10008 50	20098 50N	86421 49 •		1-2
97346 49N	97605 49N	98061 48 •		2-3
28478 48	28472 48N	80529 47 •		3-4
14331 50N	14218 50N	11578 49 •		1-3
15647 50N	15607 50	11545 49		2-4
74075 49	74145 49N	46798 48 •		1-4

69813 49			10000 50	
80255 50N	38751 50N	27869 50	91136 50	
64126 48N	48190 49	86577 49 •	12745 49	for E_1
10522 49	86797 49N	47164 49 •	11446 49	for E_2
96345 47N	38143 48	16676 49	98521 49	for E_3
99342 49	11371 49	13116 48	30786 47	for E_4
92676 49	20465 50N	89431 49 •		1-2
10758 50N	10989 50N	29582 49 •		2-3
12593 49	12594 49N	35662 48 •		3-4
13568 50N	12823 50N	21268 49 •		1-3
16027 50N	15935 50	22003 49		2-4
63452 49	63585 49N	66531 48 •		1-4

69813 49			15000 50	
93302 50N	31957 50N	16161 50	10909 51	
54992 48N	46901 49	85955 49 •	19533 49	for E_1
14599 49	86073 49N	42442 49 •	24017 49	for E_2
31866 48N	12013 49	28364 49	95085 49	for E_3
90725 49	15723 49	23895 48	55651 47	for E_4
78100 49	21284 50N	93334 49 •		1-2
10941 50N	12196 50N	60186 49 •		2-3
33391 49	33404 49N	96663 48 •		3-4
12970 50N	10249 50N	25188 49 •		1-3
16008 50N	15899 50	30321 49		2-4
53895 49	54064 49N	71854 48 •		1-4

69813 49			20000 50	
10739 51N	26685 50N	85182 49	12756 51	
61520 48N	45317 49	84969 49 •	26241 49	for E_1
17003 49	82167 49N	31315 49 •	44483 49	for E_2
78727 48N	28614 49	42276 49	85626 49	for E_3
98035 49	19392 49	34944 48	80800 47	for E_4
53652 49	22384 50N	10200 50 •		1-2
81636 49N	13334 50N	12855 50 •		2-3
70115 49	70221 49N	21158 49 •		3-4
12634 50N	55159 49N	17068 49 •		1-3
14944 50N	14850 50	34090 49		2-4
45494 49	45671 49N	69720 48 •		1-4

69813 49			25000 50	
12249 51N	31045 50N	71112 49	14642 51	
65062 48N	43557 49	83653 49 •	32597 49	for E_1
16750 49	72689 49N	13758 49 •	65164 49	for E_2
14326 49N	48090 49	52840 49	68483 49	for E_3
97323 49	22501 49	45540 48	10449 48	for E_4
28271 49	22469 50N	11016 50 •		1-2
14742 49N	14362 50N	16876 50 •		2-3
11000 50	11049 50N	35560 49 •		3-4
12004 50N	28108 49	97516 47		1-3
12392 50N	12324 50	31134 49		2-4
38265 49	38430 49N	64075 48 •		1-4

69813 49			30000 50	
13850 51N	37907 50N	10804 50	16960 51	
66602 48N	41736 49	82094 49 •	38394 49	for E_1
15532 49	63801 49N	15689 48	75403 49	for E_2
19471 49N	59623 49	56809 49	53270 49	for E_3
96619 49	25149 49	55375 48	12619 48	for E_4
10477 49	21557 50N	11529 50 •		1-2
42581 49	15258 50N	16852 50 •		2-3
13282 50	13410 50N	46607 49 •		3-4
10940 50N	38877 49	14045 49		1-3
99014 49N	98403 49	26035 49		2-4
32161 49	32308 49N	57139 48 •		1-4

69813 49			35000 50	
18531 51N	46523 50N	16800 50	18504 51	
66897 48N	39950 49	80402 49 •	43526 49	for E_1
14789 49	58774 49N	12123 49	78612 49	for E_2
23063 49N	64789 49	57853 49	43856 49	for E_3
95940 49	27418 49	64354 48	14578 48	for E_4
20902 49	20687 50N	11935 50 •		1-2
75760 49	16016 50N	15075 50 •		2-3
14312 50	14543 50N	54257 49 •		3-4
98881 49N	54175 49	19756 49		1-3
82492 49N	82022 49	22233 49		2-4
27085 49	27212 49N	50183 48 •		1-4

69813 49			40000 50	
17281 51N	55534 50N	23665 50	20468 51	
66338 48N	38265 49	78677 49 •	47975 49	for E_1
14499 49	56262 49N	19584 49	78998 49	for E_2
25766 49N	67136 49	58004 49	38143 49	for E_3
95298 49	29375 49	72482 48	16336 48	for E_4
28215 49	20061 50N	12270 50 •		1-2
95321 49	16644 50N	15075 50 •		2-3
14799 50	15146 50N	60022 49 •		3-4
89694 49N	59213 49	21618 49		1-3
71669 49N	71219 49	19633 49		2-4
22907 49	23015 49N	43734 48 •		1-4

69813 49			45000 50	
19088 51N	64498 50N	30869 50	22449 51	
65388 48N	36717 49	76991 49 •	51781 49	for E_1
14457 49	55035 49N	25176 49	78282 49	for E_2
27942 49N	68244 49	58092 49	34455 49	for E_3
94696 49	31074 49	79812 48	17906 48	for E_4
36848 49	19619 50N	12526 50 •		1-2
10793 50	17159 50N	14522 50 •		2-3
15045 50	13514 50N	64661 49 •		3-4
81769 49N	59735 49	21739 49		1-3
64023 49N	63577 49	17755 49		2-4
19405 49	19578 49N	38035 48 •		1-4

69813 49			50000 50	
20940 51N	73252 50N	38205 50	24445 51	
64230 48N	35318 49	75393 49 •	55019 49	for E_1
14539 49	54455 49N	29544 49	77138 49	for E_2
29763 49N	68754 49	58036 49	31918 49	for E_3
94135 49	32550 49	86420 48	19317 48	for E_4
45355 49	19296 50N	12708 50 •		1-2
11668 50	17579 50N	14143 50 •		2-3
15171 50	15763 50N	68532 49 •		3-4
74908 49N	58239 49	21090 49		1-3
58226 49N	57778 49	16306 49		2-4
16686 49	16766 49N	33128 48 •		1-4

69813 49			55000 50	
22830 51N	81802 50N	45506 50	26452 51	
62997 48N	34067 49	73908 49 •	57769 49	for E_1
14679 49	54206 49N	33052 49	75853 49	for E_2
31323 49N	68951 49	57963 49	30078 49	for E_3
93614 49	33863 49	92378 48	20574 48	for E_4
53193 49	19052 50N	12827 50 •		1-2
12305 50	17923 50N	13877 50 •		2-3
15230 50	15943 50N	71830 49 •		3-4
68941 49N	55871 49	20120 49		1-3
53596 49N	53143 49	15135 49		2-4
14391 49	14460 49N	28942 48 •		1-4

69813 49			60000 50	
24751 51N	90151 50N	52974 50	28469 51	
61768 48N	32954 49	72543 49 •	60110 49	for E_1
14844 49	54130 49N	35925 49	74558 49	for E_2
32680 49N	69975 49	57889 49	28490 49	for E_3
93131 49	35017 49	97763 48	21704 48	for E_4
60201 49	18862 50N	12897 50 •		1-2
12787 50	18204 50N	13685 50 •		2-3
15250 50	16082 50N	74601 49 •		3-4
63730 49N	53103 49	19044 49 •		1-3
49760 49N	49301 49	14152 49		2-4
12500 49	12561 49N	25396 48 •		1-4

69813 49			65000 50	
26696 51N	90324 50N	60350 50	30494 51	
60586 48N	31967 49	71259 49 •	62111 49	for E_1
15016 49	54144 49N	38314 49	73313 49	for E_2
33674 49N	68900 49	57819 49	27606 49	for E_3
92683 49	36044 49	10264 49	22722 48	for E_4
66382 49	18710 50N	12932 50 •		1-2
13161 50	18435 50N	13543 50 •		2-3
15246 50	16192 50N	77171 49 •		3-4
59161 49N	50447 49	17966 49		1-3
46498 49N	46033 49	13306 49		2-4
10934 49	10987 49N	22396 48 •		1-4

69813 49			70000 50	
28662 51N	10634 51N	67705 50	32526 51	
59471 48N	31089 49	70168 49 •	63831 49	for E_1
15186 49	54205 49N	40325 49	72145 49	for E_2
34934 49N	68767 49	57754 49	26739 49	for E_3
92269 49	36962 49	10707 49	23644 48	for E_4
71806 49	18585 50N	12942 50 •		1-2
13458 50	18627 50N	13435 50 •		2-3
15228 50	16282 50N	79364 49 •		3-4
55136 49N	47795 49	16933 49		1-3
43672 49N	43204 49	12565 49		2-4
96288 48	96755 48N	19854 48 •		1-4

69813 49			75000 50	
30444 51N	11423 51N	75038 50	34564 51	
58433 48N	30309 49	69143 49 •	65318 49	for E_1
15349 49	54288 49N	42037 49	71063 49	for E_2
35081 49N	68602 49	57694 49	26029 49	for E_3
91884 49	37786 49	11110 49	24476 48	for E_4
76561 49	18481 50N	12933 50 •		1-2
13697 50	18780 50N	13351 50 •		2-3
15199 50	16358 50N	81308 49 •		3-4
51575 49N	45289 49	15967 49		1-3
41190 49N	40719 49	11909 49		2-4
85324 48	85734 48N	17681 48 •		1-4

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

TABLE I(e). Ruby $\theta = 50°$.

87266 49			50000 49	
69599 50N	46507 50N	44662 50	71444 50	
36243 48N	64965 49	83184 49 •	67857 48	for E_1
65664 48	88198 49N	84802 49 •	56150 48	for E_2
32933 47N	92835 47	87535 48	99611 49	for E_3
99717 49	74794 48	61314 47	19011 47	for E_4
75117 49	20183 50N	93174 49 •	1-2	
10870 50N	10932 50N	12813 49 •	2-3	
39682 48	39731 48N	15375 48 •	3-4	
13478 50N	13312 50N	12450 49 •	1-3	
15218 50N	15148 50	13725 49	2-4	
82493 49	82631 49N	62600 48 •	1-4	

87266 49		10000 50		
83200 50N	37595 50N	32000 50	86795 50	
59874 48N	53971 49	82880 49 •	13503 49	for E_1
12599 49	82906 49N	52564 49 •	14309 49	for E_2
16858 48N	45722 48	19073 49	98043 49	for E_3
99009 49	13080 49	20181 48	66049 47	for E_4
68199 49	20562 50N	96363 49 •	1-2	
11754 50N	12223 50N	36408 49 •	2-3	
16560 49	16565 49N	67002 48 •	3-4	
12532 50N	11593 50N	21141 49 •	1-3	
15718 50N	15535 50	26692 49	2-4	
69812 49	70221 49N	88283 48 •	1-4	

87266 49		15000 50		
97949 50N	31289 50N	26111 50	10312 51	
74695 48N	52624 49	82335 49 •	19895 49	for E_1
17630 49	81881 49N	47348 49 •	27251 49	for E_2
47869 48N	12334 49	31062 49	94127 49	for E_3
98032 49	19338 49	37587 48	12101 48	for E_4
56484 49	21769 50N	10229 50 •	1-2	
11452 50N	13406 50N	74878 49 •	2-3	
39060 49	39110 49N	16277 49 •	3-4	
11710 50N	89166 49N	23041 49 •	1-3	
15681 50N	15532 50	36808 49	2-4	
58534 49	59077 49N	94188 48 •	1-4	

87266 49		20000 50		
11370 51N	28675 50N	22120 50	12025 51	
84143 48N	51097 49	61574 49 •	25765 49	for E_1
20909 49	76949 49N	37853 49 •	43553 49	for E_2
10082 49N	24116 49	43378 49	86232 49	for E_3
96909 49	23967 49	55753 48	17675 48	for E_4
41700 49	21913 50N	11016 50 •	1-2	
88112 49N	14448 50N	12158 50 •	2-3	
69451 49	59747 49N	30379 49 •	3-4	
10988 50N	52156 49N	16414 49 •	1-3	
14773 50N	14517 50	41666 49	2-4	
48541 49	49446 49N	90341 48 •	1-4	

87266 49		25000 50		
13031 51N	30272 50N	22554 50	13802 51	
86908 48N	49519 49	80656 49 •	31008 49	for E_1
22117 49	74000 49N	25539 49 •	58156 49	for E_2
16321 49N	35975 49	52807 49	75171 49	for E_3
95726 49	27982 49	73208 48	22915 48	for E_4
29607 49	21964 50N	11731 50 •	1-2	
41089 49N	15336 50N	15512 50 •	2-3	
98123 49	99151 49N	45078 49 •	3-4	
10198 50N	14648 49N	51200 48 •	1-3	
13073 50N	12837 50	40756 49	2-4	
40779 49	41349 49N	82332 48 •	1-4	

87266 49		30000 50		
14764 51N	35002 50N	26401 50	15632 51	
93264 48N	47976 49	79649 49 •	35598 49	for E_1
22176 49	69127 49N	14156 49 •	67298 49	for E_2
21953 49N	44121 49	56090 49	64775 49	for E_3
94547 49	31191 49	89536 48	27661 48	for E_4
24558 49	21507 50N	12251 50 •	1-2	
22558 48	16077 50N	16766 50 •	2-3	
11736 50	11971 50N	59083 49 •	3-4	
93446 49N	11962 49	44069 48	1-3	
11267 50N	11052 50	37024 49	2-4	
34136 49	34663 49N	73119 48 •	1-4	

87266 49		35000 50		
14555 51N	41619 50N	32149 50	17502 51	
95067 48N	46521 49	78611 49 •	39567 49	for E_1
22026 49	65645 49N	52947 48 •	71954 49	for E_2
26439 49N	48690 49	60691 49	56980 49	for E_3
93410 49	33995 49	10426 49	31876 48	for E_4
25240 49	20941 50N	12639 50 •	1-2	
34212 49	16687 50N	16922 50 •	2-3	
12926 50	13239 50N	69314 49 •	3-4	
85395 49N	27149 49	10264 49	1-3	
97826 49N	95803 49	33079 49	2-4	
28726 49	29200 49N	64132 48 •	1-4	

87266 49		40000 50		
18396 51N	48924 50N	38811 50	19407 51	
95880 48N	45179 49	77587 49 •	42978 49	for E_1
21969 49	63450 49N	13786 48	74090 49	for E_2
29993 49N	51080 49	61970 49	51485 49	for E_3
92935 49	36380 49	11745 49	35583 48	for E_4
28894 49	20446 50N	12944 50 •	1-2	
55922 49	17187 50N	16747 50 •	2-3	
13417 50	14033 50N	77273 49 •	3-4	
78281 49N	34926 49	13403 49	1-3	
86377 49N	84425 49	29702 49	2-4	
24332 49	24754 49N	55981 48 •	1-4	

87266 49		45000 50		
20275 51N	56536 50N	45897 50	21339 51	
96074 48N	43959 49	76604 49 •	45903 49	for E_1
22026 49	62115 49N	64921 48	74929 49	for E_2
32874 49N	52279 49	62631 49	47575 49	for E_3
91333 49	38421 49	12919 49	38028 48	for E_4
33724 49	20047 50N	13186 50 •	1-2	
70927 49	17595 50N	16520 50 •	2-3	
13732 50	14563 50N	83623 49 •	3-4	
72009 49N	30570 49	14919 49	1-3	
77462 49N	75550 49	26919 49	2-4	
20762 49	21135 49N	48846 48 •	1-4	

87266 49		50000 50		
22186 51N	64244 50N	53176 50	23293 51	
95885 48N	42859 49	75677 49 •	48415 49	for E_1
22159 49	61308 49N	10515 49	75097 49	for E_2
35764 49N	52822 49	62986 49	44709 49	for E_3
90407 49	40176 49	13964 49	41671 48	for E_4
38820 49	19730 50N	13376 50 •	1-2	
81669 49	17929 50N	16317 50 •	2-3	
13890 50	14938 50N	88806 49 •	3-4	
66694 49N	39923 49	15511 49	1-3	
70332 49N	68445 49	24612 49	2-4	
17849 49	18178 49N	42714 48 •	1-4	

87266 49		55000 50		
24123 51N	71957 50N	60535 50	25265 51	
95468 48N	41871 49	74815 49 •	50580 49	for E_1
22335 49	60822 49N	13758 49	74916 49	for E_2
37281 49N	52998 49	63178 49	42540 49	for E_3
89554 49	41696 49	14894 49	44167 48	for E_4
43761 49	19476 50N	13524 50 •	1-2	
89626 49	18205 50N	16151 50 •	2-3	
13956 50	15217 50N	93110 49 •	3-4	
61971 49N	40008 49	15584 49	1-3	
64477 49N	62609 49	22672 49	2-4	
15460 49	15750 49N	37489 48 •	1-4	

87266 49		60000 50		
26001 51N	79635 50N	67921 50	27252 51	
94920 48N	40984 49	74017 49 •	52455 49	for E_1
22533 49	60533 49N	16425 49	74552 49	for E_2
39008 49N	52962 49	63200 49	40852 49	for E_3
88771 49	43021 49	15722 49	46361 48	for E_4
48369 49	19269 50N	13638 50 •	1-2	
95693 49	18432 50N	16019 50 •	2-3	
13967 50	15433 50N	96732 49 •	3-4	
57812 49N	39391 49	15366 49	1-3	
59564 49N	57713 49	21015 49	2-4	
13486 49	13744 49N	33046 48 •	1-4	

87266 49		65000 50		
28056 51N	87263 50N	75306 50	29252 51	
94304 48N	40189 49	73283 49 •	54086 49	for E_1
22737 49	60368 49N	18656 49	74099 49	for E_2
40504 49N	52805 49	63327 49	39504 49	for E_3
88053 49	44182 49	16464 49	46303 48	for E_4
52586 49	19098 50N	13725 50 •	1-2	
10043 50	18622 50N	15913 50 •	2-3	
13945 50	15607 50N	99816 49 •	3-4	
54131 49N	38396 49	14988 49	1-3	
55369 49N	53536 49	19583 49	2-4	
11845 49	12074 49N	29270 48 •	1-4	

87266 49		70000 50		
30046 51N	94836 50N	82675 50	31362 51	
93659 48N	39474 49	72609 49 •	55315 49	for E_1
22941 49	60279 49N	20546 49	73606 49	for E_2
41810 49N	52577 49	63342 49	38407 49	for E_3
87394 49	45207 49	17129 49	50026 48	for E_4
56410 49	18954 50N	13790 50 •	1-2	
10421 50	18782 50N	15829 50 •	2-3	
13901 50	15749 50N	10246 50 •	3-4	
50856 49N	37207 49	14528 49	1-3	
51738 49N	49924 49	18330 49	2-4	
10471 49	10675 49N	26050 48 •	1-4	

87266 49		75000 50		
32048 51N	10235 51N	90024 50	33281 51	
93010 48N	38829 49	71990 49 •	56772 49	for E_1
23138 49	60240 49N	22160 49	73104 49	for E_2
42961 49N	52311 49	63337 49	37497 49	for E_3
86789 49	46116 49	17727 49	51563 48	for E_4
59662 49	18832 50N	13839 50 •	1-2	
10727 50	10916 50N	15761 50 •	2-3	
13845 50	15868 50N	10476 50 •	3-4	
47927 49N	35935 49	14032 49	1-3	
48558 49N	46765 49	17225 49	2-4	
93113 48	94946 48N	23297 48 •	1-4	

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

TABLE I(f). Ruby $\theta = 60°$.

10471 50			50000 49	
70754 50N	45768 50N	47857 50	68664 50	
45080 48N	59603 49	79794 49 •	70995 48	for E_1
73510 48	79767 49N	59480 49 •	67200 48	for E_2
57507 47N	10940 48	97054 48	99520 49	for E_3
99625 49	85927 48	83456 47	40402 47	for E_4
53524 49	20164 50N	97495 49 •		1-2
11725 50N	11849 50N	15290 49 •		2-3
49949 48	49938 48N	29120 48 •		3-4
12719 50N	12489 50N	12721 49 •		1-3
14842 50N	14711 50	15531 49		2-4
09144 49	09610 49N	77651 48 •		1-4

10471 50			10000 50	
85708 50N	36726 50N	40781 50	81653 50	
74914 48N	58690 49	79459 49 •	13619 49	for E_1
14414 49	79188 49N	57050 49 •	16330 49	for E_2
27022 48N	48354 48	20572 49	97704 49	for E_3
98634 49	16163 49	28560 48	33241 48	for E_4
47874 49	20620 50N	10171 50 •		1-2
12391 50N	13192 50N	41127 49 •		2-3
19753 49	19768 49N	11685 49 •		3-4
11652 50N	10536 50N	20127 49 •		1-3
15472 50N	15110 50	30841 49		2-4
74701 49	75816 49N	10945 49 •		1-4

10471 50			15000 50	
10183 51N	30864 50N	36666 50	96034 50	
94620 48N	57435 49	78963 49 •	19401 49	for E_1
20428 49	77810 49N	51983 49 •	28737 47	for E_2
68578 48N	11366 49	32126 49	93763 49	for E_3
97191 49	22748 49	55058 48	24644 48	for E_4
39471 49	21200 50N	10826 50 •		1-2
11698 50N	14361 50N	77183 49 •		2-3
42612 49	42771 49N	26111 49 •		3-4
10717 50N	7 221 49N	20730 49 •		1-3
15463 50N	14915 50	42630 49		2-4
61 36 49	63384 49N	11609 49 •		1-4

10471 50			20000 50	
13880 51N	20669 50N	35960 50	11159 51	
10768 49N	56100 49	78363 49 •	24413 49	for E_1
24661 49	75323 49N	44240 49 •	41964 49	for E_2
12801 49N	19328 49	42809 49	87349 49	for E_3
95456 49	28381 49	83310 48	36265 48	for E_4
30761 49	21585 50N	11562 50 •		1-2
92652 49N	15398 50N	11505 50 •		2-3
60478 49	69244 49N	44343 49 •		3-4
98761 49N	49499 49N	15447 49 •		1-3
14679 50N	14038 50	48366 49		2-4
50928 49	52664 49N	11062 49 •		1-4

10471 50			25000 50	
13664 51N	29889 50N	38408 50	13813 51	
11638 49N	54790 49	77709 49 •	28701 49	for E_1
27059 49	72160 49N	35348 49 •	53019 49	for E_2
19348 49N	26336 49	50882 49	79642 49	for E_3
93583 49	33124 49	11076 49	47058 48	for E_4
24708 49	21580 50N	12210 50 •		1-2
57561 49N	16214 50N	14340 50 •		2-3
90558 49	92797 49N	62478 49 •		3-4
90747 49N	22750 49N	78502 48 •		1-3
13355 50N	12687 50	48498 49		2-4
41960 49	43708 49N	99715 48 •		1-4

10471 50			30000 50	
15497 51N	33588 50N	45103 50	14545 51	
12223 49N	53557 49	77042 49 •	32350 49	for E_1
28273 49	69208 49N	27145 49 •	60613 49	for E_2
25360 49N	31095 49	56037 49	72439 49	for E_3
91695 49	37079 49	13603 49	56576 48	for E_4
22349 49	21296 50N	12717 50 •		1-2
24054 49N	16879 50N	15953 50 •		2-3
10569 50	11035 50N	78153 49 •		3-4
83230 49N	31739 48N	11614 48 •		1-3
11909 50N	11245 50	45518 49		2-4
34721 49	36378 49N	87769 48 •		1-4

10471 50			35000 50	
17374 51N	38796 50N	49110 50	16342 51	
12618 49N	52425 49	76309 49 •	35453 49	for E_1
28944 49	66923 49N	20390 49 •	65327 49	for E_2
30419 49N	33821 49	59130 49	66583 49	for E_3
89875 49	40360 49	15857 49	64749 48	for E_4
22817 49	20919 50N	13110 50 •		1-2
26690 48	17409 50N	16740 50 •		2-3
11476 50	12258 50N	90601 49 •		3-4
76462 49N	94968 48	36039 48		1-3
10597 50N	99416 49	41575 49		2-4
28924 49	30446 49N	76363 48 •		1-4

10471 50			40000 50	
19285 51N	44853 50N	55806 50	18189 51	
12886 49N	51398 49	75766 49 •	38090 49	for E_1
29407 49	65299 49N	15026 49 •	69157 49	for E_2
34567 49N	35183 49	60954 49	82062 49	for E_3
88171 49	43084 49	17835 49	71660 48	for E_4
24897 49	20553 50N	13421 50 •		1-2
22642 49	17831 50N	17121 50 •		2-3
11975 50	13110 50N	10049 50 •		3-4
70502 49N	17227 49	66960 48		1-3
94829 49N	88403 49	37679 49		2-4
24290 49	25664 49N	66187 48 •		1-4

10471 50			45000 50	
21224 51N	51380 50N	62847 50	20077 51	
13067 49N	50474 49	75181 49 •	40365 49	for E_1
29792 49	64178 49N	10768 49 •	69840 49	for E_2
37963 49N	35727 49	62043 49	58591 49	for E_3
86604 49	45354 49	19557 49	77495 48	for E_4
27728 49	20230 50N	13674 50 •		1-2
37429 49	18168 50N	17293 50 •		2-3
12219 50	13723 50N	10818 50 •		3-4
65283 49N	21782 49	86093 48		1-3
85524 49N	79234 49	34165 49		2-4
20576 49	21804 49N	57402 48 •		1-4

10471 50			50000 50	
23185 51N	58170 50N	70051 50	21997 51	
13190 49N	49644 49	74636 49 •	42318 49	for E_1
30141 49	63407 49N	73423 48 •	70830 49	for E_2
40766 49N	35798 49	62707 49	55895 49	for E_3
85179 49	47258 49	21052 49	82403 48	for E_4
30817 49	19955 50N	13880 50 •		1-2
48504 49	10438 50N	17373 50 •		2-3
12311 50	14181 50N	11427 50 •		3-4
60703 49N	24353 49	97453 48		1-3
77734 49N	71587 49	31093 49		2-4
17583 49	18676 49N	49938 48 •		1-4

10471 50			55000 50	
25164 51N	65108 50N	77326 50	23942 51	
13271 49N	48899 49	74133 49 •	44011 49	for E_1
30469 49	62876 49N	45416 48 •	71397 49	for E_2
43103 49N	35603 49	63118 49	53763 49	for E_3
83890 49	48865 49	22352 49	86553 48	for E_4
33904 49	19723 50N	14051 50 •		1-2
56949 49	10657 50N	17407 50 •		2-3
12313 50	14533 50N	11914 50 •		3-4
56668 49N	25690 49	10377 49		1-3
71158 49N	65156 49	28432 49		2-4
15152 49	16125 49N	43637 48 •		1-4

10471 50			60000 50	
27158 51N	72130 50N	84624 50	25908 51	
13324 49N	48229 49	73669 49 •	45488 49	for E_1
30777 49	62506 49N	22159 48 •	71699 49	for E_2
45071 49N	35262 49	63374 49	52046 49	for E_3
82727 49	50234 49	23485 49	90080 48	for E_4
36861 49	19526 50N	14192 50 •		1-2
63509 49	10836 50N	17420 50 •		2-3
12260 50	14813 50N	12309 50 •		3-4
53094 49N	26253 49	10685 49		1-3
65550 49N	59696 49	26120 49		2-4
13161 49	14029 49N	38526 48 •		1-4

10471 50			65000 50	
29164 51N	79199 50N	91920 50	27891 51	
13357 49N	47625 49	73242 49 •	46705 49	for E_1
31060 49	62247 49N	25694 47 •	71832 49	for E_2
46744 49N	34846 49	63531 49	50640 49	for E_3
81677 49	51408 49	24478 49	93097 48	for E_4
39632 49	19357 50N	14310 50 •		1-2
68696 49	10983 50N	17482 50 •		2-3
12177 50	13040 50N	12638 50 •		3-4
49912 49N	26330 49	10781 49		1-3
60722 49N	53016 49	24187 49		2-4
11519 49	12294 49N	33840 48 •		1-4

10471 50			70000 50	
31180 51N	86293 50N	99204 50	29888 51	
13375 49N	47080 49	72830 49 •	47931 49	for E_1
31341 49	62065 49N	14138 48	71850 49	for E_2
46180 49N	34396 49	63624 49	49471 49	for E_3
80730 49	52423 49	25352 49	95694 48	for E_4
42199 49	19212 50N	14409 50 •		1-2
72861 49	19104 50N	17419 50 •		2-3
12076 50	15229 50N	12899 50 •		3-4
47067 49N	26102 49	10741 49		1-3
56527 49N	50970 49	22381 49		2-4
10151 49	10846 49N	30038 48 •		1-4

10471 50			75000 50	
33204 51N	93399 50N	10647 51	31897 51	
13383 49N	46586 49	72488 49 •	48948 49	for E_1
31596 49	61987 49N	28543 48	71814 49	for E_2
49423 49N	33938 49	63676 49	48483 49	for E_3
79873 49	53307 49	26125 49	97944 48	for E_4
44562 49	19085 50N	14492 50 •		1-2
76254 49	19207 50N	17414 50 •		2-3
11968 50	15387 50N	13122 50 •		3-4
44513 49N	25683 49	10613 49		1-3
52853 49N	47442 49	20848 49		2-4
90039 48	96286 48N	28801 48 •		1-4

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

TABLE I(g). Ruby $\theta = 70°$.

12217 50			50000 49	
71636 50N	45223 50N	51319 50	65539 50	
53100 48N	63657 49	76606 49 •	71413 48	for E_1
79208 48	76525 49N	63435 49 •	75513 48	for E_2
10063 48N	11645 48	10308 49	99455 49	for E_3
99539 49	94972 48	10656 48	78833 47	for E_4
34391 49	20197 50N	10018 50 •	1-2	
12348 50N	12609 50N	17121 49 •	2-3	
58015 48	58014 48N	53564 48 •	3-4	
12056 50N	11732 50N	12508 49 •	1-3	
14503 50N	14265 50	16082 49	2-4	
94767 49	95892 49N	91636 48 •	1-4	

12217 50			10000 50	
87606 50N	36133 50N	47925 50	75814 50	
89479 48N	62747 49	76204 49 •	13248 49	for E_1
15925 49	73576 49N	61054 49 •	17520 49	for E_2
43619 48N	65507 48	21214 49	97520 49	for E_3
98220 49	10172 49	36993 48	26968 48	for E_4
30449 49	20651 50N	10510 50 •	1-2	
12734 50N	14052 50N	43829 49 •	2-3	
22011 49	22065 49N	20743 49 •	3-4	
10960 50N	95991 49N	18368 49 •	1-3	
15331 50N	14584 50	34156 49	2-4	
78464 49	81429 49N	13063 49 •	1-4	

12217 50			15000 50	
10474 51N	30606 50N	47388 50	87959 50	
11443 49N	61662 49	75699 49 •	18340 49	for E_1
22962 49	73745 49N	56522 49 •	28975 49	for E_2
10102 49N	93084 48	31834 49	93796 49	for E_3
96123 49	25936 49	78359 48	51193 48	for E_4
25139 49	21129 50N	11208 50 •	1-2	
11671 50N	15343 50N	77363 49 •	2-3	
44484 49	44990 49N	43617 49 •	3-4	
10033 50N	70797 49N	18301 49 •	1-3	
15441 50N	14191 50	47069 49	2-4	
63694 49	68073 49N	13092 49 •	1-4	

12217 50			20000 50	
12271 51N	28664 50N	49588 50	10179 51	
13181 49N	60565 49	75164 49 •	22546 49	for E_1
28225 49	71229 49N	50377 49 •	39896 49	for E_2
17345 49N	13880 49	40790 49	88558 49	for E_3
93427 49	32641 49	12167 49	75691 48	for E_4
20313 49	21371 50N	11915 50 •	1-2	
92995 49N	16419 50N	11002 50 •	2-3	
66709 49	98860 49N	69165 49 •	3-4	
92184 49N	45919 49N	13970 49 •	1-3	
14773 50N	13172 50	52943 49	2-4	
51154 49	56241 49N	13135 49 •	1-4	

12217 50			25000 50	
14129 51N	29688 50N	53907 50	11707 51	
14417 49N	59531 49	74638 49 •	26022 49	for E_1
31710 49	68586 49N	43873 49 •	48635 49	for E_2
24731 49N	17068 49	47272 49	82839 49	for E_3
90415 49	38216 49	16415 49	97509 48	for E_4
17321 49	21329 50N	12526 50 •	1-2	
84196 49N	17264 50N	13536 50 •	2-3	
83633 49	89198 49N	92552 49 •	3-4	
84745 49N	25396 49N	85275 48 •	1-3	
13605 50N	11817 50	52776 49	2-4	
40975 49	46183 49N	11715 49 •	1-4	

12217 50			30000 50	
16032 51N	32786 50N	59568 50	13354 51	
15315 49N	58588 49	76142 49 •	28910 49	for E_1
33938 49	66281 49N	37986 49 •	54881 49	for E_2
31309 49N	18695 49	51491 49	77580 49	for E_3
87368 49	42716 49	20211 49	11551 49	for E_4
16296 49	21110 50N	13018 50 •	1-2	
37625 49N	17900 50N	15244 50 •	2-3	
94135 49	13474 50N	11137 50 •	3-4	
77961 49N	10540 49N	37699 48 •	1-3	
12282 50N	10425 50	49300 49	2-4	
32956 49	37920 49N	10149 49 •	1-4	

12217 50			35000 50	
17969 51N	37216 50N	65969 50	15094 51	
15982 49N	57739 49	73682 49 •	31329 49	for E_1
35423 49	64463 49N	33054 49 •	59135 49	for E_2
36769 49N	19168 49	34106 49	73164 49	for E_3
86485 49	46294 49	23468 49	12974 49	for E_4
16688 49	20820 50N	13414 50 •	1-2	
16083 49N	10360 50N	16334 50 •	2-3	
99560 49	11621 50N	12549 50 •	3-4	
71860 49N	53806 47N	20101 47 •	1-3	
11017 50N	91615 49	44678 49	2-4	
24733 49	31278 49N	86806 48 •	1-4	

12217 50			40000 50	
19932 51N	42478 50N	72746 50	16905 51	
16487 49N	56982 49	73262 49 •	33373 49	for E_1
36488 49	63089 49N	29034 49 •	62010 49	for E_2
61169 49N	18951 49	55706 49	69589 49	for E_3
81864 49	49127 49	26200 49	14079 49	for E_4
17911 49	20543 50N	13735 50 •	1-2	
50454 47	18712 50N	17030 50 •	2-3	
10167 50	12470 50N	13578 50 •	3-4	
66431 49N	59698 48	23010 48	1-3	
98937 49N	80753 49	40034 49	2-4	
21924 49	25996 49N	74003 48 •	1-4	

12217 50			45000 50	
21915 51N	48264 50N	79705 50	19771 51	
16877 49N	56307 49	72879 49 •	35116 49	for E_1
37306 49	62063 49N	25766 49 •	63971 49	for E_2
44687 49N	18371 49	56692 49	66719 49	for E_3
79538 49	51300 49	28472 49	14931 49	for E_4
19549 49	20280 50N	14000 50 •	1-2	
13110 49	18966 50N	17405 50 •	2-3	
10182 50	13111 50N	14322 50 •	3-4	
61625 49N	10132 49	40000 48	1-3	
69237 49N	71626 49	35810 49	2-4	
18194 49	21803 49N	63222 48 •	1-4	

12217 50			50000 50	
23914 51N	54389 50N	66748 50	20670 51	
17184 49N	55705 49	72531 49 •	36616 49	for E_1
37970 49	61293 49N	23089 49 •	65331 49	for E_2
47510 49N	17627 49	57307 49	64404 49	for E_3
77496 49	53190 49	30363 49	15593 49	for E_4
21350 49	20046 50N	14223 50 •	1-2	
22731 49	19156 50N	17792 50 •	2-3	
10090 50	13608 50N	14860 50 •	3-4	
57375 49N	12757 49	51316 48	1-3	
80953 49N	64001 49	32123 49	2-4	
15276 49	18462 49N	54274 48 •	1-4	

12217 50			55000 50	
25926 51N	60740 50N	93823 50	22618 51	
17428 49N	55167 49	72215 49 •	37916 49	for E_1
38529 49	60708 49N	20870 49 •	66289 49	for E_2
49794 49N	16827 49	57696 49	62517 49	for E_3
75711 49	54660 49	31943 49	16110 49	for E_4
23174 49	19840 50N	14411 50 •	1-2	
30157 49	19300 50N	18005 50 •	2-3	
99438 49	14002 50N	15254 50 •	3-4	
53609 49N	14375 49	58709 48	1-3	
73882 49N	57618 49	28957 49	2-4	
12970 49	15783 49N	46881 48 •	1-4	

12217 50			60000 50	
27948 51N	67245 50N	10090 51	24582 51	
17625 49N	54685 49	71928 49 •	39053 49	for E_1
39011 49	60259 49N	19008 49 •	66974 49	for E_2
51661 49N	16030 49	57946 49	60961 49	for E_3
74151 49	55869 49	33274 49	16519 49	for E_4
24946 49	19661 50N	14572 50 •	1-2	
35970 49	19412 50N	18159 50 •	2-3	
97727 49	14321 50N	15546 50 •	3-4	
50263 49N	15331 49	63398 48	1-3	
67823 49N	52243 49	26250 49	2-4	
11126 49	13614 49N	40767 48 •	1-4	

12217 50			65000 50	
29979 51N	73860 50N	10799 51	26566 51	
17786 49N	54230 49	71666 49 •	40054 49	for E_1
39435 49	59908 49N	17428 49 •	67469 49	for E_2
53205 49N	15265 49	58106 49	59663 49	for E_3
72784 49	56875 49	34405 49	16846 49	for E_4
28629 49	19503 50N	14710 50 •	1-2	
40586 49	19499 50N	18272 50 •	2-3	
95942 49	14505 50N	15764 50 •	3-4	
47280 49N	15846 49	66220 48	1-3	
62602 49N	47682 49	23932 49	2-4	
96352 48	11842 49N	35690 48 •	1-4	

12217 50			70000 50	
32017 51N	80553 50N	11507 51	28565 51	
17920 49N	53837 49	71428 49 •	40941 49	for E_1
39812 49	59630 49N	16071 49 •	67829 49	for E_2
54496 49N	14546 49	58210 49	58567 49	for E_3
71581 49	57722 49	35373 49	17113 49	for E_4
28208 49	19364 50N	14831 50 •	1-2	
44302 49	19569 50N	18358 50 •	2-3	
94170 49	14806 50N	15928 50 •	3-4	
44608 49N	16067 49	67747 48	1-3	
58073 49N	43781 49	21940 49	2-4	
84153 48	10382 49N	31450 48 •	1-4	

12217 50			75000 50	
34061 51N	87303 50N	12214 51	30577 51	
18031 49N	53501 49	71209 49 •	41732 49	for E_1
40151 49	59408 49N	14896 49 •	69092 49	for E_2
55585 49N	13876 49	58276 49	57632 49	for E_3
70519 49	58443 49	36209 49	17328 49	for E_4
29678 49	19241 50N	14936 50 •	1-2	
47333 49	19625 50N	18424 50 •	2-3	
92484 49	14994 50N	16054 50 •	3-4	
42206 49N	16090 49	68372 48	1-3	
54117 49N	40419 49	20218 49	2-4	
74072 48	91669 48N	27865 48 •	1-4	

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

TABLE I(h). Ruby $\theta = 80°$.

13962 50			50000 49	
72188 50N	44891 50N	54929 50	62149 50	
60029 48N	67167 49	73513 49 •	69495 48	for E_1
82494 48	73376 49N	66951 49 •	80821 48	for E_2
21352 48N	10775 48	10556 49	99412 49	for E_3
99455 49	10163 49	13428 48	18832 48	for E_4
16819 49	22202 50N	10167 50 •	1-2	
12787 50N	13336 50N	18355 49 •	2-3	
63116 48	63168 48N	12019 49 •	3-4	
13530 50N	10971 50N	11857 49 •	1-3	
14252 50N	13711 50	17598 49	2-4	
99271 49	10257 50N	10489 49 •	1-4	

13962 50			10000 50	
88785 50N	35790 50N	55159 50	69417 50	
10300 49N	56387 49	73007 49 •	12514 49	for E_1
17073 49	72009 49N	64809 49 •	17968 49	for E_2
85619 48N	33651 48	20971 49	97341 49	for E_3
97616 49	19897 49	54707 48	67156 48	for E_4
14805 49	20565 50N	10714 50 •	1-2	
12565 50N	15128 50N	45704 49 •	2-3	
23151 49	23483 49N	45158 49 •	3-4	
10609 50N	85607 49N	16340 49 •	1-3	
15537 50N	13530 50	35597 49	2-4	
79600 49	89157 49N	15484 49 •	1-4	

13962 50			15000 50	
10653 51N	30467 50N	57824 50	79176 50	
13378 49N	65495 49	72426 49 •	16903 49	for E_1
25179 49	69605 49N	60985 49 •	28518 49	for E_2
19117 49N	52670 48	29938 49	93529 49	for E_3
94121 49	29944 49	11782 49	12630 49	for E_4
12254 49	21081 50N	11422 50 •	1-2	
10807 50N	16827 50N	78550 49 •	2-3	
43816 49	46192 49N	89160 49 •	3-4	
98751 49N	59917 49N	15025 49 •	1-3	
16032 50N	12556 50	47235 49	2-4	
60926 49	75663 49N	16791 49 •	1-4	

13962 50			20000 50	
12506 51N	28665 50N	62308 50	91417 50	
15520 49N	64627 49	71861 49 •	20377 49	for E_1
31566 49	66661 49N	56196 49 •	37440 49	for E_2
28323 49N	56691 48	36368 49	88560 49	for E_3
89204 49	36706 49	18849 49	18438 49	for E_4
10106 49	21250 50N	12114 50 •	1-2	
81673 49N	19150 50N	10963 50 •	2-3	
60116 49	58627 49N	12955 50 •	3-4	
91771 49N	37738 49N	11068 49 •	1-3	
15503 50N	10901 50	49889 49	2-4	
45234 49	62267 49N	15806 49 •	1-4	

13962 50			25000 50	
14412 51N	29588 50N	67866 50	10584 51	
17292 49N	63834 49	71342 49 •	23158 49	for E_1
36131 49	63809 49N	51407 49 •	44499 49	for E_2
37091 49N	47393 48	40340 49	83513 49	for E_3
83783 49	42790 49	25302 49	22560 49	for E_4
88671 48	21188 50N	12706 50 •	1-2	
55661 49N	19008 50N	13398 50 •	2-3	
60673 49	87130 49N	15793 50 •	3-4	
84705 49N	21522 49N	69736 48 •	1-3	
14269 50N	91490 49	46401 49	2-4	
33338 49	50119 49N	13772 49 •	1-4	

13962 50			30000 50	
16357 51N	32346 50N	73970 50	12179 51	
18569 49N	63129 49	70879 49 •	25415 49	for E_1
39303 49	61884 49N	47179 49 •	49611 49	for E_2
43762 49N	31456 48	42631 49	79103 49	for E_3
78709 49	47295 49	30540 49	25204 49	for E_4
84931 48	20995 50N	13191 50 •	1-2	
34643 49N	19503 50N	15115 50 •	2-3	
71141 49	10177 50N	17468 50 •	3-4	
77042 49N	10755 49N	37321 48 •	1-3	
12796 50N	76174 49	40671 49	2-4	
24867 49	40057 49N	11562 49 •	1-4	

13962 50			35000 50	
18331 51N	36412 50N	00344 50	13937 51	
19566 49N	62506 49	70470 49 •	27275 49	for E_1
41548 49	59441 49N	43651 49 •	53244 49	for E_2
48598 49N	13951 48	43961 49	75521 49	for E_3
74357 49	50574 49	34579 49	26786 49	for E_4
87215 48	20755 50N	13587 50 •	1-2	
18998 49N	19771 50N	16284 50 •	2-3	
70786 49	11270 50N	18359 50 •	3-4	
71509 49N	38821 48N	14159 48 •	1-3	
11378 50N	63850 49	34929 49	2-4	
18954 49	32164 49N	95835 48 •	1-4	

13962 50			40000 50	
20325 51N	41243 50N	86061 50	15763 51	
20362 49N	61959 49	70110 49 •	28827 49	for E_1
43194 49	57912 49N	40762 49 •	55845 49	for E_2
52080 49N	26987 47N	44772 49	72683 49	for E_3
70761 49	52981 49	37661 49	27701 49	for E_4
93038 48	20511 50N	13914 50 •	1-2	
76090 48N	19914 50N	17085 50 •	2-3	
68010 49	12103 50N	18796 50 •	3-4	
65843 49N	47170 47	17871 47	1-3	
10128 50N	54204 49	29954 49	2-4	
14796 49	26107 49N	79507 48 •	1-4	

13962 50			45000 50	
22335 51N	46609 50N	93465 50	17649 51	
21007 49N	61476 49	69791 49 •	30139 49	for E_1
44449 49	56709 49N	38397 49 •	57740 49	for E_2
54618 49N	17664 48N	45302 49	70437 49	for E_3
67021 49	54705 49	40030 49	28218 49	for E_4
10067 49	20281 50N	14188 50 •	1-2	
72370 47	19909 50N	17644 50 •	2-3	
65300 49	12750 50N	18985 50 •	3-4	
60346 49N	32399 48	12643 48	1-3	
90646 49N	46666 49	25961 49	2-4	
11816 49	21460 49N	66386 48 •	1-4	

13962 50			50000 50	
24358 51N	52345 50N	10012 51	19579 51	
21539 49N	61050 49	69509 49 •	31261 49	for E_1
45438 49	55752 49N	36446 49 •	59149 49	for E_2
56503 49N	30825 48N	45672 49	68642 49	for E_3
65412 49	56169 49	41881 49	28500 49	for E_4
10906 49	20973 50N	14420 50 •	1-2	
69162 48	20028 50N	19046 50 •	2-3	
62602 49	13261 50N	19042 50 •	3-4	
56458 49N	50078 48	20016 48	1-3	
81686 49N	40717 49	22542 49	2-4	
96284 48	17889 49N	55935 48 •	1-4	

13962 50			55000 50	
26390 51N	58341 50N	10682 51	21541 51	
21984 49N	60671 49	69259 49 •	32229 49	for E_1
46240 49	54981 49N	34819 49 •	60219 49	for E_2
57933 49N	42321 48N	45945 49	67189 49	for E_3
63420 49	57254 49	43352 49	28642 49	for E_4
11760 49	19887 50N	14619 50 •	1-2	
11606 49	20048 50N	18342 50 •	2-3	
60093 49	13674 50N	19030 50 •	3-4	
52602 49N	61348 48	25006 48	1-3	
74138 49N	35958 49	19049 49	2-4	
79848 48	15094 49N	47590 48 •	1-4	

13962 50			60000 50	
28431 51N	64526 50N	11356 51	23527 51	
22361 49N	60334 49	69035 49 •	33073 49	for E_1
46906 49	54351 49N	33448 49 •	61047 49	for E_2
59040 49N	52366 48N	44165 49	65996 49	for E_3
61756 49	58122 49	44543 49	28701 49	for E_4
12594 49	19720 50N	14792 50 •	1-2	
15226 49	20057 50N	18566 50 •	2-3	
57827 49	14013 50N	18981 50 •	3-4	
49201 49N	69442 48	28358 48	1-3	
67746 49N	32093 49	17648 49	2-4	
67229 48	12883 49N	40877 48 •	1-4	

13962 50			65000 50	
30478 51N	70855 50N	12032 51	25530 51	
22683 49N	60031 49	68834 49 •	33814 49	for E_1
47470 49	53830 49N	32279 49 •	61700 49	for E_2
59913 49N	61174 48N	46341 49	65002 49	for E_3
60352 49	58830 49	45522 49	28709 49	for E_4
13392 49	19372 50N	14943 50 •	1-2	
18070 49	20059 50N	18739 50 •	2-3	
55808 49	14296 50N	18915 50 •	3-4	
46189 49N	72765 48	30569 48	1-3	
62791 49N	28912 49	15831 49	2-4	
57350 48	11110 49N	35429 48 •	1-4	

13962 50			70000 50	
32530 51N	77287 50N	12711 51	27547 51	
22942 49N	59759 49	68633 49 •	34471 49	for E_1
47953 49	53393 49N	31273 49 •	62222 49	for E_2
60613 49N	60935 48N	46490 49	64165 49	for E_3
59153 49	59417 49	46339 49	28687 49	for E_4
14145 49	19440 50N	15075 50 •	1-2	
20342 49	20059 50N	18875 50 •	2-3	
54016 49	14535 50N	18842 50 •	3-4	
43508 49N	75223 48	31979 48	1-3	
57600 49N	26257 49	14315 49	2-4	
49484 48	96715 48N	30961 48 •	1-4	

13962 50			75000 50	
34587 51N	83805 50N	13391 51	29576 51	
23205 49N	59512 49	68489 49 •	35035 49	for E_1
48374 49	53022 49N	30399 49 •	62643 49	for E_2
61181 49N	75811 48N	46618 49	63430 49	for E_3
58121 49	59911 49	47029 49	28648 49	for E_4
14850 49	19322 50N	15193 50 •	1-2	
22183 49	20057 50N	18984 50 •	2-3	
52426 49	14739 50N	18767 50 •	3-4	
41109 49N	76411 48	32819 48	1-3	
53532 49N	24015 49	13038 49	2-4	
43122 48	84891 49N	27263 48 •	1-4	

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

TABLE I(i). Ruby $\theta = 90°$.

15707 50			50000 49	
72375 50N	44779 50N	58425 50	58729 50	
65592 48N	70405 49	70405 49 •	65592 48	for E_1
83105 48	70220 49N	70220 49 •	83105 48	for E_2
70412 49N	65600 48N	65584 48	70398 49	for E_3
70228 49	83112 48	83098 48	70212 49	for E_4
20000 42N	20207 50N	10214 50 •	1-2	
18873 46	19157 50N	25894 49 •	2-3	
75774 45N	64937 48N	29772 50 •	3-4	
15175 50N	14703 46	14979 45	1-3	
19176 50N	21057 46	28597 45	2-4	
16656 46	14974 50N	15933 49 •	1-4	

15707 50			10000 50	
89185 50N	35678 50N	61285 50	63578 50	
11507 49N	69767 49	69767 49 •	11507 49	for E_1
17784 49	68437 49N	68437 49 •	17784 49	for E_2
69768 49N	11508 49N	11507 49	69767 49	for E_3
68438 49	17784 49	17784 49	68436 49	for E_4
43000 43	20669 50N	10777 50 •	1-2	
17040 45	20399 50N	58696 49 •	2-3	
15520 45N	23806 49N	29057 50 •	3-4	
13191 50N	10520 45	29580 44	1-3	
19997 50N	29810 45	79800 44	2-4	
19460 45	12286 50N	22438 49 •	1-4	

15707 50			15000 50	
10713 51N	30423 50N	65288 50	72273 50	
15219 49N	69053 49	69053 49 •	15219 49	for E_1
27000 49	65352 49N	65352 49 •	27000 49	for E_2
69053 49N	15219 49N	15219 49	69053 49	for E_3
65352 49	27000 49	27000 49	65352 49	for E_4
90000 42	21064 50N	11491 50 •	1-2	
56000 44N	21035 50N	91975 49 •	2-3	
53000 44	46569 49N	27898 50 •	3-4	
11511 50N	26500 44N	60000 43 •	1-3	
19327 50N	27200 44N	12500 44 •	2-4	
17200 44N	95985 49N	22389 49 •	1-4	

15707 50			20000 50	
12585 51N	28664 50N	70050 50	84464 50	
18039 49N	68370 49	68370 49 •	18039 49	for E_1
34649 49	61639 49N	61639 49 •	34649 49	for E_2
68370 49N	18039 49N	18039 49	68370 49	for E_3
61639 49	34649 49	34649 49	61639 49	for E_4
23000 43	21211 50N	12179 50 •	1-2	
33000 43N	21211 50N	11990 50 •	2-3	
43000 43N	68549 49N	26536 50 •	3-4	
10132 50N	40000 42N	30000 43	1-3	
17545 50N	27600 44	16100 46	2-4	
14700 44	72880 49N	19336 49 •	1-4	

15707 50			25000 50	
14507 51N	29558 50N	75329 50	99308 50	
20223 49N	67756 49	67756 49 •	20223 49	for E_1
40372 49	58052 49N	58052 49 •	40372 49	for E_2
67756 49N	20223 49N	20223 49	67756 49	for E_3
58052 49	40372 49	40372 49	58052 49	for E_4
31000 43N	21142 50N	12765 50 •	1-2	
84000 43N	21150 50N	14064 50 •	2-3	
16900 44	86749 49N	25233 50 •	3-4	
90056 49N	49000 43N	82000 43 •	1-3	
15402 50N	38000 43N	42000 43 •	2-4	
22000 43N	55122 49N	15732 49 •	1-4	

15707 50			30000 50	
16467 51N	32264 50N	80970 50	11596 51	
21950 49N	67217 49	67217 49 •	21950 49	for E_1
44484 49	54965 49N	54965 49 •	44484 49	for E_2
67217 49N	21950 49N	21950 49	67217 49	for E_3
54965 49	44484 49	44484 49	54964 49	for E_4
22000 43N	20957 50N	13247 50 •	1-2	
26000 43N	21007 50N	15527 50 •	2-3	
58000 43	10084 50N	24120 50 •	3-4	
80805 49N	62000 43N	82000 43 •	1-3	
13390 50N	73000 43	35000 43	2-4	
22000 43	42205 49N	12587 49 •	1-4	

15707 50			35000 50	
18452 51N	36165 50N	86873 50	13381 51	
23341 49N	66746 49	66746 49 •	23341 49	for E_1
47441 49	52433 49N	52433 49 •	47441 49	for E_2
66746 49N	23341 49N	23341 49	66746 49	for E_3
52433 49	47441 49	47441 49	52433 49	for E_4
25000 43N	20728 50N	13643 50 •	1-2	
24000 43N	20855 50N	16551 50 •	2-3	
10700 44	11159 50N	23213 50 •	3-4	
73137 49N	44000 43N	65000 43 •	1-3	
11677 50N	16000 43N	35000 43 •	2-4	
28000 43N	32933 49N	10102 49 •	1-4	

15707 50			40000 50	
20457 51N	40860 50N	92970 50	15246 51	
24482 49N	66336 49	66336 49 •	24482 49	for E_1
49610 49	50386 49N	50386 49 •	49610 49	for E_2
66336 49N	24482 49N	24482 49	66336 49	for E_3
50386 49	49610 49	49610 49	50306 49	for E_4
25000 43	20496 50N	13972 50 •	1-2	
80000 42N	20720 50N	17278 50 •	2-3	
73000 43	11985 50N	22484 50 •	3-4	
66711 49N	70000 42	42000 43	1-3	
10267 50N	14000 43	20000 42 •	2-4	
12000 43N	26222 49N	81964 48 •	1-4	

15707 50			45000 50	
22476 51N	46092 50N	99212 50	17164 51	
25433 49N	65978 49	65978 49 •	25433 49	for E_1
51241 49	48726 49N	48726 49 •	51241 49	for E_2
65978 49N	25433 49N	25433 49	65978 49	for E_3
48726 49	51241 49	51241 49	48726 49	for E_4
34000 43	20270 50N	14249 50 •	1-2	
10000 42	20608 50N	17806 50 •	2-3	
38000 43	12631 50N	21895 50 •	3-4	
61267 49N	26000 43	56000 43	1-3	
91162 49N	58000 43	29000 43	2-4	
50000 42	21279 49N	67394 48 •	1-4	

15707 50			50000 50	
24506 51N	51700 50N	10556 51	19120 51	
26235 49N	65663 49	65663 49 •	26235 49	for E_1
52500 49	47367 49N	47367 49 •	52500 49	for E_2
65663 49N	26235 49N	26235 49	65663 49	for E_3
47367 49	52500 49	52500 49	47367 49	for E_4
20000 42	20078 50N	14484 50 •	1-2	
60000 42	20516 50N	18198 50 •	2-3	
	13146 50N	21416 50 •	3-4	
56608 49N	80000 42	60000 42 •	1-3	
81717 49N	11300 44	88000 43	2-4	
35000 43	17565 49N	56165 48 •	1-4	

15707 50			55000 50	
26546 51N	57578 50N	11201 51	21102 51	
26921 49N	65385 49	65385 49 •	26921 49	for E_1
53493 49	46242 49N	46242 49 •	53493 49	for E_2
65385 49N	26921 49N	26921 49	65385 49	for E_3
46242 49	53493 49	53493 49	46242 49	for E_4
21000 43	19898 50N	14687 50 •	1-2	
60000 42	20442 50N	18496 50 •	2-3	
55000 43	13564 50N	21021 50 •	3-4	
52583 49N	20000 42	11000 43	1-3	
73897 49N	21000 43	40000 42	2-4	
50000 42N	14719 49N	47402 48 •	1-4	

15707 50			60000 50	
28593 51N	63656 50N	11853 51	23105 51	
27512 49N	65138 49	65138 49 •	27512 49	for E_1
54294 49	45300 49N	45300 49 •	54294 49	for E_2
65138 49N	27512 49N	27512 49	65138 49	for E_3
45300 49	54294 49	54294 49	45300 49	for E_4
40000 42N	19737 50N	14864 50 •	1-2	
90000 42	20381 50N	18727 50 •	2-3	
25000 43	13909 50N	20692 50 •	3-4	
49075 49N	19000 43N	24000 43 •	1-3	
67351 49N	38000 43	27000 43	2-4	
12000 43	12498 49N	40469 48 •	1-4	

15707 50			65000 50	
30645 51N	69885 50N	12510 51	25183 51	
28028 49N	64918 49	64918 49 •	28028 49	for E_1
54951 49	44500 49N	44500 49 •	54951 49	for E_2
64918 49N	28028 49N	28028 49	64918 49	for E_3
44500 49	54951 49	54951 49	44500 49	for E_4
18000 43N	19592 50N	15019 50 •	1-2	
13000 43	20332 50N	18909 50 •	2-3	
62000 43	14197 50N	20414 50 •	3-4	
45993 49N	17000 43N	41000 43 •	1-3	
61812 49N	11000 43N	10000 43 •	2-4	
11000 43	10735 49N	34910 48 •	1-4	

15707 50			70000 50	
32703 51N	76231 50N	13173 51	27153 51	
28481 49N	64720 49	64720 49 •	28481 49	for E_1
55498 49	43816 49N	43816 49 •	55498 49	for E_2
64720 49N	28481 49N	28481 49	64720 49	for E_3
43816 49	55498 49	55498 49	43816 49	for E_4
15000 43N	19463 50N	15155 50 •	1-2	
12000 43	20291 50N	19055 50 •	2-3	
24000 43	14442 50N	20176 50 •	3-4	
43266 49N	26000 43N	44000 43 •	1-3	
57077 49N	20000 43	15000 43	2-4	
10000 43	93160 48N	30396 48 •	1-4	

15707 50			75000 50	
34764 51N	82668 50N	13839 51	29191 51	
28883 49N	64542 49	64542 49 •	28883 49	for E_1
55961 49	43223 49N	43223 49 •	55961 49	for E_2
64542 49N	28883 49N	28883 49	64542 49	for E_3
43223 49	55961 49	55961 49	43223 49	for E_4
32000 43	19346 50N	15277 50 •	1-2	
14000 43	20257 50N	19174 50 •	2-3	
12000 43	14652 50N	19971 50 •	3-4	
40838 49N	10000 43	29000 43	1-3	
52989 49N	56000 43	57000 43	2-4	
80000 42	81567 48N	26606 48 •	1-4	

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

θ in radians			H_{dc} in kilogauss	
E_1	E_2	E_3	E_4	E_j in Kmc/s
a_1	b_1	c_1	d_1	for E_1
a_2	b_2	c_2	d_2	for E_2
a_3	b_3	c_3	d_3	for E_3
a_4	b_4	c_4	d_4	for E_4
α	β	γ	1-2 transition	
α	β	γ	2-3 transition	
α	β	γ	3-4 transition	
α	β	γ	1-3 transition	
α	β	γ	2-4 transition	
α	β	γ	1-4 transition	

FIG. 23. Format of each sub-block in tables for ruby.

The energy eigenvectors $|j\rangle$ and the energy eigenvalues E_j are defined by the equations

$$\mathcal{H}|j\rangle = E_j|j\rangle$$

and

$$|j\rangle = a_j|\tfrac{3}{2}\rangle + b_j|\tfrac{1}{2}\rangle + c_j|-\tfrac{1}{2}\rangle + d_j|-\tfrac{3}{2}\rangle,$$

where $j=1$ through 4 denotes the four energy levels.

The axes were chosen so that the c axis of the ruby crystal is the z axis, the dc magnetic field H_{dc} lies in the x–z plane at an angle θ from the z axis, and the y axis is perpendicular to the plane containing H_{dc} and the c axis.

The transition probability matrix elements between levels k and l are given by

$$\frac{(\bar{H}')_{kl}}{\beta_M H_{rf}} = \frac{1}{\beta_M H_{rf}}\langle k|2\beta_M \mathbf{H}_{rf}\cdot\mathbf{S}|l\rangle$$

$$= (\alpha\phi_1 + \gamma\phi_3) + j\beta\phi_2,$$

where H_{rf} is the peak amplitude of the linearly-polarized rf magnetic field, and ϕ_1, ϕ_2, ϕ_3 are the direction cosines of $\mathbf{H}_{rf}$ with respect to the x, y, z axes.

For each operating point, characterized by H_{dc} and θ, the four energy eigenvalues E_j in units of kmc/s, the eigenvector expansion coefficients $a_j \cdots d_j$ for each level j, and the values of α, β, γ for each possible transition are presented in tabular form in Tables I(a) through I(i). The tables cover $\theta=10°$ to $\theta=90°$ in 10° increments, with H_{dc} increasing from 500 gauss to 7500 gauss by 500 gauss increments in each table.

The format of each subblock in the tables is shown in Fig. 23. The numbers in the blocks are expressed in

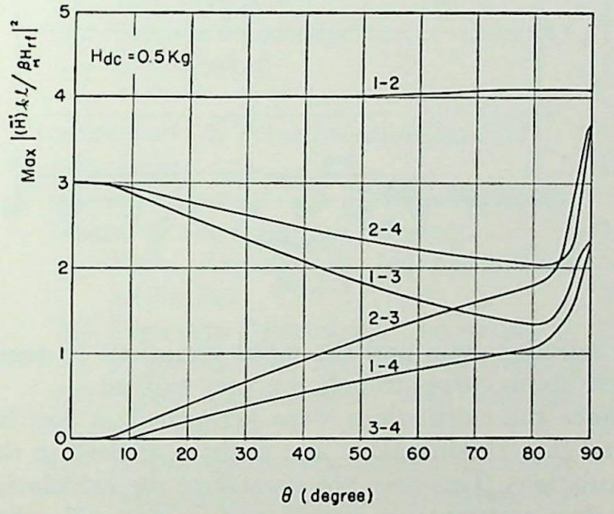


FIG. 24.

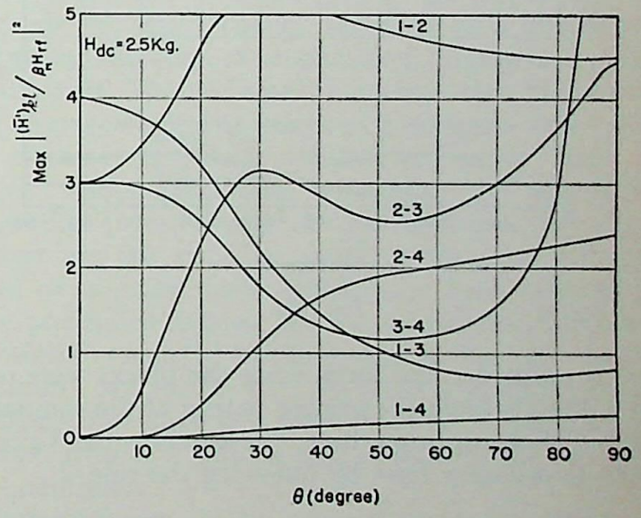


FIG. 26.

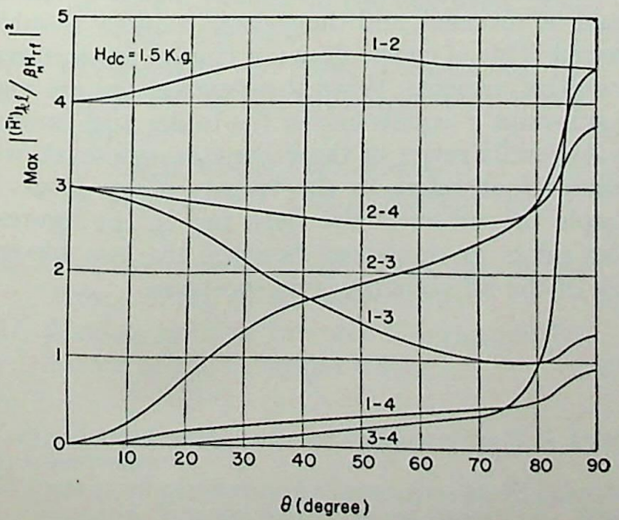


FIG. 25.

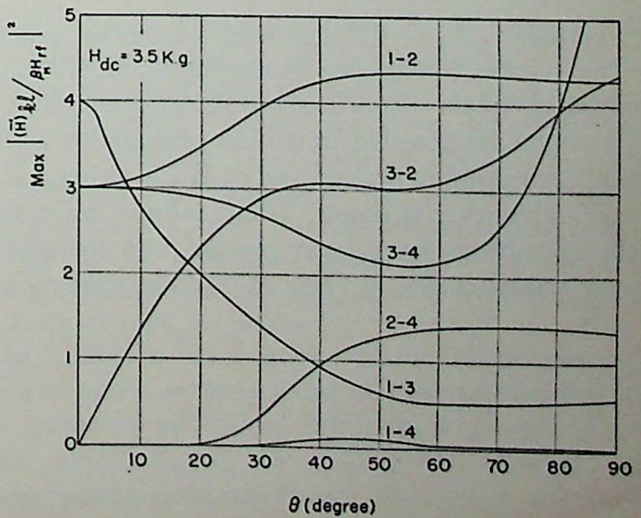


FIG. 27.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

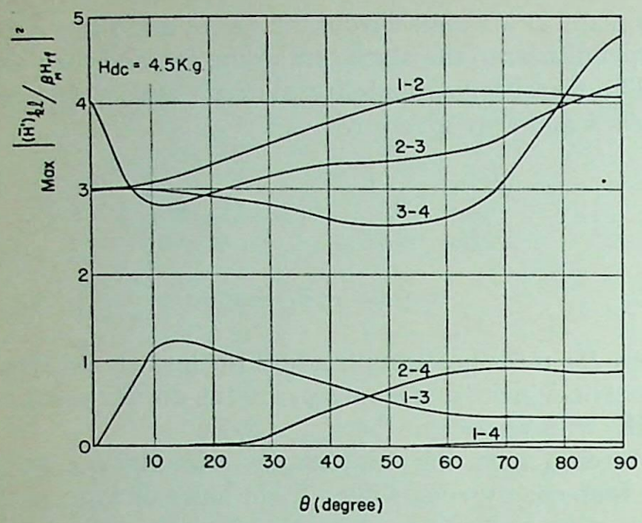


FIG. 28.

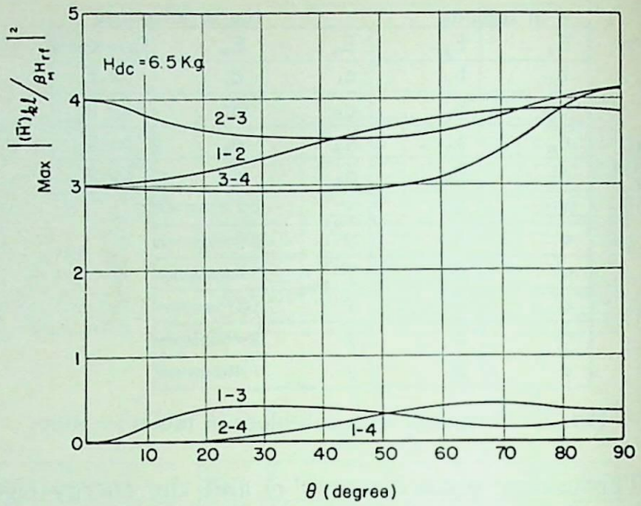


FIG. 30.

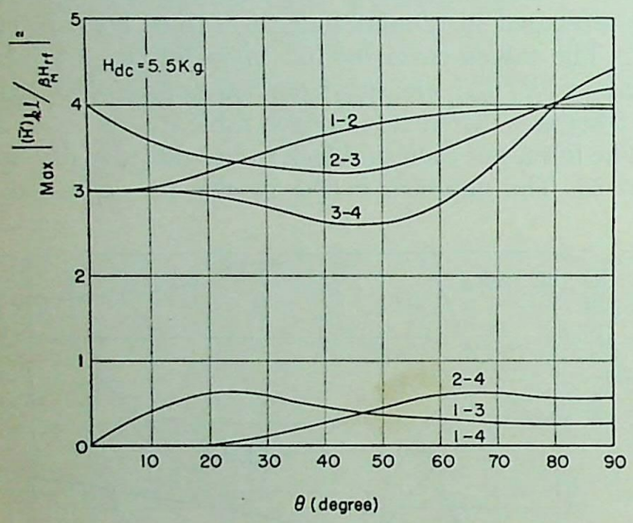


FIG. 29.

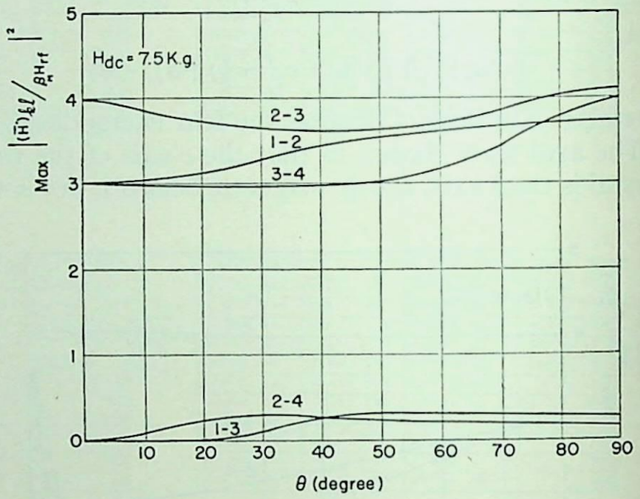


FIG. 31.

floating-point decimal form, since the blocks were prepared directly from the printed output of the computer to avoid transcription errors. The numbers can be converted to ordinary form by following the rule

$$abcde\ gh \equiv a \cdot bcde \times 10^{(gh\ 50)}.$$

The symbols *, N, or CR following a number indicate that the number is negative. For example, 56 423 51CR in the tables equals -56.423 in ordinary decimal form.

The maximum possible transition probability matrix element for optimum orientation of $\mathbf{H}_{rf}$ is either $\beta_M{}^2$ or $(\alpha^2+\gamma^2)$, whichever is larger. The maximum values for all the transitions have been plotted *vs* θ for several different values of H_{dc} in Figs. 24–31. The operation of the selection rules and the rapid variations in regions where levels curve strongly are very evident.

Since the calculations were performed, it has been found that D is negative, not positive as used in these calculations. This does not invalidate the calculations, but does require some simple modifications. To obtain correct answers for negative D, the signs of the E_j's should be reversed, and the signs of α and γ should be reversed. This of course does not change the squares of the matrix elements. When these corrections are made, the subscript J equals one in the tables and curves of this Appendix refers to the highest energy level, while J equals four refers to the lowest energy level. For example the subscript one dash two in the figures or tables refers to transitions between the two top-most levels of the ruby energy level spectrum.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

Polarization Measurements on Nuclear Gamma Rays*

LAWRENCE W. FAGG,† *Nucleonics Division, U. S. Naval Research Laboratory, Washington, D. C.*

AND

STANLEY S. HANNA,† *Physics Division, Argonne National Laboratory, Lemont, Illinois*

TABLE OF CONTENTS

I. Theoretical Survey 711
A. Introduction 711
B. Linear Polarization Distribution 712
C. Circular Polarization Distribution 717

II. Polarization-Sensitive Processes 719
A. Compton Effect 719
(a) Sensitivity to Plane Polarization 719
(b) Sensitivity to Circular Polarization . . 720
B. Photodisintegration of the Deuteron 722
C. Photoelectric Effect 724
D. Other Polarization-Sensitive Processes . . . 725
(a) Pair Production 725
(b) Nuclear Photoeffect 726
(c) Photofission 726

III. Measurements of Polarization of Nuclear Gamma Rays 726
A. Measurement of Plane Polarization by Means of the Compton Effect 726
(a) Early Experiments on Annihilation Radiation 727
(b) Direction-Polarization Correlation . . . 727
(c) Experiments of Metzger and Deutsch 728
(d) Other Measurements with Radioactive Sources 730
(e) Measurements on Gamma Rays from Nuclear Reactions and Coulomb Excitation 733
(f) Measurements on Gamma Rays from Aligned Nuclei 734
B. Measurement of Plane Polarization by Means of the Photodisintegration of the Deuteron 735
(a) Experimental Considerations 735
(b) Experiments Using Nuclear Emulsions 736
C. Measurement of Plane Polarization by Means of the Photoelectric Effect 737
D. Measurement of Circular Polarization by Means of the Compton Effect 737
(a) Experimental Considerations 738
(b) Measurements on Gamma Rays from Nuclei (Parity Conserving Reactions) 739
(c) Beta-Gamma Circular Polarization Correlation 741
E. Tabulation of Polarization Experiments on Nuclear Gamma Rays 744
Bibliography 756

I. THEORETICAL SURVEY

A. Introduction

THE investigation of gamma radiation has encompassed all the important properties of electromagnetic radiation: energy, intensity as a function of energy and direction of propagation, multipolarity, and polarization. The emission of gamma rays of discrete and characteristic energies was associated early with the existence of sharp, well-defined energy levels in nuclei, whose properties determine the intensities (transition probabilities) of the radiations. Of equal interest are the radiation patterns (angular distributions) of emitting nuclei which have been oriented in some prescribed fashion, for the spatial distribution is intimately connected to the multipole character of the radiation (My 38, Ha 40). By relatively simple measurements, therefore, it is often possible to establish the multipolarity of a gamma ray, at least in the case of a pure multipole.

There still remains an essential ambiguity, since the radiation may be either electric or magnetic. This property is associated with the parity of the radiation. The ambiguity in the parity is immediately resolved if the linear polarization of the radiation is known (Fa 48, Ha 48). Moreover, a knowledge of the polarization can be used to untangle the results in the case of mixed multipoles (Zi 50). It is sometimes possible to resolve such ambiguities with indirect arguments based on observations of associated phenomena. The measurement of polarization, however, provides a direct means of attack on such problems.

The radiation from nuclei may also be circularly polarized and a measurement of this polarization provides valuable additional information (Go 46, Fa 49, Ha 51, To 52, St 53). In a discussion of linear and circular polarization two types of nuclear orientation are distinguished, *alignment* and *polarization*. In the

* Work performed partly under the auspices of the U. S. Atomic Energy Commission.

† LWF now at Atlantic Research Corporation, 901 N. Columbus Street, Alexandria, Virginia. SSH—Visiting Guggenheim Fellow (1958–1959), Clarendon Laboratory, Oxford, England.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

TABLE I. Summary of conditions which produce nuclear alignment or polarization, and the corresponding types of correlations. The abbreviations dir, pol, circ, and align stand for direction, (linear) polarization, circular polarization, and alignment, respectively. The symbol LT(B) represents a low temperature orientation of the Bleaney type which produces alignment, LT(RG) represents a method of the Rose-Gorter type which leads to polarization. In each case only the most complex condition or correlation is listed. Thus if the circular polarization of the initial gamma ray is measured, its direction also is measured and one may have a dir-dir or dir-pol correlation as well as a circ-circ correlation.

Initial particle	Initial condition	Type of orientation	Type of correlation
p, n, etc.	dir	align	p-γ dir-pol
p, n, etc.	pol	pol	p-γ pol-circ
β	dir	pol	β-γ dir-circ
γ	dir	align	γ-γ dir-pol
γ	pol	align	γ-γ pol-pol
γ	circ	pol	γ-γ circ-circ
...	LT(B)	align	"align"-pol
...	LT(RG)	pol	"pol"-circ

case of *alignment* an axis of rotational symmetry is established, but the sense of this axis is not determined. In the case of *polarization* an axis is established which provides a single preferred direction in space. If the spin of the nuclear state is j and its component along the axis is m and if the (relative) population of the mth sublevel is $w(m)$, then the condition for alignment is $w(m)=w(-m)$ and that for polarization is $w(m) \neq w(-m)$. If the $w(m)$ are all equal, the nuclei are not oriented.

From an assemblage of aligned nuclei the radiation can show linear polarization but not circular polarization (Ha 48). Only if there is nuclear polarization will there be a component of circular polarization. Moreover, the processes which produce nuclear orientation can be classified according to the type of orientation which is produced, either alignment or polarization. In Table I the various processes are summarized and classified in this manner and the possible polarization correlation for each case is listed (in addition to the references already cited, see Si 51, Wo 49, To 52a, Si 53a, Le 56). This tabulation illustrates the formal way of designating a correlation. For example, a γ-γ direction-polarization correlation represents the measurement made on a γ-γ cascade in which the (linear) polarization of one gamma ray is determined relative to the direction of the other gamma ray. Two general methods are available for producing orientation. In one a nuclear particle is observed with or without polarization, and in the other an external or atomic field is employed at very low temperature.

The literature on the theory of direction-direction correlations is voluminous and detailed (see the comprehensive list of references given in De 57a). In the present survey, the theoretical results are presented (without derivation) in a manner which stresses the application to polarization experiments in order to provide a basis for the discussion in Part III.

B. Linear Polarization Distribution

The general expression for the intensity distribution of linearly polarized radiation has been given in several different forms by many authors (Ha 48, Zi 50, Ll 52, St 53, To 53, Sa 53, Si 53a, Bi 53, Sa 55b, Ma 58). The various expressions are equivalent and differ only superficially to cover different experimental situations or degrees of complexity. In this review we give three forms, two of which are generally appropriate for the direction-polarization correlation but differ enough to make one or the other preferable in certain problems, while the other is designed specifically for the case of nuclear orientation at low temperature. It is convenient to make such a classification, partly for historical reasons, although in many instances the differences are hardly more than a matter of nomenclature.

The direction-direction correlation for two radiations in a nuclear process is given in a concise form by (Co 53, Bi 53, De 57a)

$$W(\theta)=\sum_{\nu} A_{\nu}(1)A_{\nu}(2)P_{\nu}(\cos\theta). \qquad \text{(I-1)}$$

Here $P_{\nu}(\cos\theta)$ is the ordinary Legendre polynomial, θ is the angle between the directions of the two radiations, and 1 and 2 stand for sets of quantum numbers for the first and second radiations, respectively. The summation over ν is over all integral values (including zero) for which the coefficient of $P_{\nu}(\cos\theta)$ does not vanish. If we let $1 \rightarrow L_1 L_1' jj'$ and $2 \rightarrow L_2 L_2' jj'$ and sum the expression over all allowed sets of values of these quantum numbers, we can treat problems in which both the intermediate state and the radiations are of a mixed nature. The quantities L_n and L_n' represent possible values of the total angular momentum (multipolarity) of the nth radiation, while j and j' are possible total angular momenta for the intermediate state. Because of their great complexity, problems of this general type have not been studied extensively, especially in connection with polarization studies of gamma rays.

In most of the cases which have been studied experimentally, the intermediate state is characterized by unique spin and parity and the radiations are mixtures involving not more than two components. With these restrictions Eq. (I-1) is greatly simplified, not so much in form as in the complexity of the computation. We adhere to the notation which seems to provide the greatest ease in calculation through the use of published tables. The summation in Eq. (I-1) is now restricted to the even integers (including zero) and, after setting $j=j'$ for the intermediate state, the factorization of the coefficient of $P_{\nu}(\cos\theta)$ is complete. Hence, writing out the summation over the L's explicitly, we have (Bi 53, Fe 55)

$$A_{\nu}(n)=F_{\nu}(L_n L_n j_n j)+2\delta_n F_{\nu}(L_n L_n' j_n j) + \delta_n{}^2 F_{\nu}(L_n' L_n' j_n j), \qquad \text{(I-2)}$$

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

which is the expression that is applicable to gamma radiation. The intensity ratio of the L_n and L_n' radiation is denoted by δ_n^2 (δ_n is real) and j_n is the total angular momentum of the nth state, i.e., either the initial (1) or the final (2) state. Equation (I-1), is normalized so that $A_0(n)=1+\delta_n^2$. The most complete compilation of the functions $F_\nu(L_nL_n'j_nj)$ is that given in Fe 55 which covers all cases of apparent interest for even ν.

If the nth radiation is a material particle (e.g., an alpha particle), then it is only necessary to replace each $F_\nu(L_nL_n'j_nj)$ in Eq. (I-2) by $F_\nu(L_nL_n'j_nj)b_\nu(L_nL_n')$ and δ_n by $\delta_n\cos\xi$ (leaving δ_n^2 unchanged, however), where the "particle factor" (Bi 53) is

$$b_\nu(LL')=\frac{2[L(L+1)L'(L'+1)]^{\frac{1}{2}}}{L(L+1)+L'(L'+1)-\nu(\nu+1)}, \quad \text{(I-3)}$$

and ξ is an appropriate phase angle. If, in addition, the particle has a spin, it is convenient to employ the channel spin formalism. The total correlation function is then an incoherent superposition of the functions corresponding to each value of the channel spin obtained by compounding the spins of the two interacting particles. In beta decay and in Coulomb excitation the particle factors are more complex in that they involve specific features of the interactions. These factors are discussed in Al 56 (Coulomb excitation) and Bi 53 (beta decay). The β-γ circular polarization correlation is discussed in Sec. C following.

If unobserved intervening radiations are present, (I-1) is generalized (Sa 54), for transitions between nuclear states of well-defined spin and parity, to

$$W(\theta)=\sum_\nu A_\nu(1)U_\nu(2)\cdots U_\nu(n-1)A_\nu(n)P_\nu(\cos\theta). \quad \text{(I-1')}$$

The function

$$U_\nu(m)=(-1)^{j_{m+1}-j_m-L_m}[(2j_{m+1}+1)(2j_m+1)]^{\frac{1}{2}} \times W(j_mj_mj_{m+1}j_{m+1};\nu L_m)$$

is inserted for each unobserved transition $j_m(L_m)j_{m+1}$. The quantity $W(j_mj_mj_{m+1}j_{m+1};\nu L_m)$ is the Racah coefficient. For mixed radiation $U_\nu(L_m)+\delta_m^2U_\nu(L_m')$ (normalized to $1+\delta_m^2$) is used. The intervening radiation may be any kind of particle.

TABLE II(a). Numerical values of $P_\nu^{(2)}(\cos\theta)$.

$\cos\theta$ \ ν	2	3	4	5	6
0	3.00	0	−7.50	0	13.12
0.1	2.97	1.49	−6.90	−5.05	10.70
0.2	2.88	2.88	−5.18	−8.87	4.19
0.3	2.73	4.10	−2.52	−10.46	−4.21
0.4	2.52	5.04	0.76	−9.18	−11.41
0.5	2.25	5.62	4.22	−4.92	−14.15
0.6	1.92	5.76	7.30	1.61	−10.11
0.7	1.53	5.35	9.30	8.81	0.69
0.8	1.08	4.32	9.40	13.90	14.16
0.9	0.57	2.56	6.65	12.82	20.13
1.0	0	0	0	0	0

TABLE II(b). Numerical values of the coefficients $\kappa_\nu(LL')$.

LL' \ ν	2	3	4	5	8
11	−1/2				
12	−1/6	−1/6			
13	−1/12	−1/12	−1/12		
22	1/2	0	−1/12		
23	−1/4	1/4	−1/60	−1/20	
33	1/3	0	1/3	0	−1/30

The prescription for obtaining the linear polarization distribution in a direction-polarization experiment can now be stated simply as follows. In each term of (I-1) characterized by (L_2L_2') and ν, replace $P_\nu(\cos\theta)$ by the expression (Ll 52, Sa 55b, De 57a),

$$P_\nu(\cos\theta)+(\pm)_{L_2'}\cos2\gamma\kappa_\nu(L_2L_2')P_\nu^{(2)}(\cos\theta). \quad \text{(I-4)}$$

Without loss of generality it is assumed that the linear polarization is observed for the second radiation. The direction of the linear polarization is specified by γ, the angle between the polarization vector and the plane of the reaction. In practice only two values of γ (0 and 90°) are usually considered, so that $\cos2\gamma=\pm1$. In the discussion of the experiments, the intensities of polarized radiation for these two directions is designated by J_0 and J_{90} or $J_{\parallel}$ and $J_\perp$ (J_θ, J_ϕ are also used in the literature). The quantity often given is the ratio J_0/J_{90}, although theoretically the polarization excess or difference (J_0-J_{90}) is the more natural quantity, as pointed out by Fano (Fa 49). Thus we write

$$P_1W(\theta)=(J_0-J_{90}) =\sum(\pm)_{L_2'}A_\nu\kappa_\nu(L_2L_2')P_\nu^{(2)}(\cos\theta), \quad \text{(I-5)}$$

which may be normalized to the total unpolarized intensity

$$W(\theta)=(J_0+J_{90})=\sum A_\nu P_\nu(\cos\theta) \quad \text{(I-6)}$$

if so desired. In these expressions the coefficient $A_\nu(1)A_\nu(2)$ or $A_\nu(1)U_\nu(2)\cdots A_\nu(n)$ has been shortened to A_ν and the sum is over ν and whatever quantum numbers are appropriate. The $P_\nu^{(2)}(\cos\theta)$ appearing in the polarization distribution is the unnormalized associated Legendre function. Numerical values for $P_\nu^{(2)}(\cos\theta)$ are given in Table II(a). The other quantity $\kappa_\nu(LL')$ appearing in (I-4) depends on the multipolarity of the radiation. It is given by (Ll 52)

$$\kappa_\nu(LL')=-\left[\frac{(\nu-2)!}{(\nu+2)!}\right]^{\frac{1}{2}}\frac{C(LL'\nu,11)}{C(LL'\nu,1-1)}, \quad \text{(I-7)}$$

where $C(LL'\nu,11)$ is the familiar vector addition coefficient. Numerical values for these coefficients are presented in Table II(b). Finally, in (I-4) the + or −

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

TABLE III(a). Numerical values of the coefficients in the series $a_\nu(LL)=\Sigma_n \alpha_n A_n(LL)$. For odd ν, all $\alpha_n=0$.

L	ν	α_2	α_4	α_6
1	0	−1		
	2	1		
2	0	1	−1/6	
	2	−1	−5/6	
	4	0	1	
3	0	2/3	2/3	−1/15
	2	−2/3	10/3	−1/3
	4	0	−4	−9/15
	6	0	0	1

sign is to be employed according as the $2^{L_2'}$-pole radiation is electric or magnetic in character.‡

For the case of mixed radiation from a single intermediate level the polarization distribution can be written more explicitly as

$$W(\theta,\gamma)=W(\theta)$$
$$+(\pm)_{L_2'}\cos 2\gamma \sum_\nu A_\nu(1)\mathcal{A}_\nu(2)P_\nu^{(2)}(\cos\theta). \quad \text{(I-8)}$$

The sum is over even ν (excluding zero). The coefficient $A_\nu(1)$ is given by (I-2) or a suitable modification, if the first radiation is not a gamma ray or there are intervening radiations. The new coefficient $\mathcal{A}_\nu(2)$ differs only in the inclusion of the κ's with appropriate signs [so as to leave unchanged the sign convention in (I-8)]. Thus

$$\mathcal{A}_\nu(2)=-\kappa_\nu(L_2L_2)F_\nu(L_2L_2j_2j)$$
$$+2\delta_2\kappa_\nu(L_2L_2')F_\nu(L_2L_2'j_2j)$$
$$+\delta_2{}^2\kappa_\nu(L_2'L_2')F_\nu(L_2'L_2'j_2j). \quad \text{(I-9)}$$

In this case $L_2'=L_2+1$.

In a number of the problems studied experimentally, the complexity of the direction-direction correlation is limited to $\nu=0$ and 2. The direction-polarization correlation then becomes [with $\mathcal{A}_2=A_2(1)\mathcal{A}_2(2)$]

$$W=A_0+A_2P_2\pm\cos 2\gamma\,\mathcal{A}_2P_2^{(2)}. \quad \text{(I-10)}$$

Unless the directional correlation is nearly isotropic, a measurement of the polarization easily determines the correct sign in this expression and hence the parity of the radiation. The measurement can also be decisive in removing an ambiguity regarding the mixture of the radiation, as can be seen by inspecting (I-9) and Table II(b). There is one interesting exception in this particular example, namely, the case of pure dipole or pure quadrupole radiation. For both $L=1$ and $L=2$ we have, using (I-9) and Table II(b)

$$W=1+A_2P_2\pm\tfrac{1}{2}\cos 2\gamma A_2P_2^{(2)} \quad \text{(I-10')}$$

and for $\theta=90°$, we have

$$\frac{J_0}{J_{90}}=\frac{1-\frac{1}{2}A_2\pm\frac{3}{2}A_2}{1-\frac{1}{2}A_2\mp\frac{3}{2}A_2}. \quad \text{(I-10'')}$$

However, this expression does not hold for a dipole-quadrupole mixture. It is a quite general result that the polarization measurement can provide vital information on mixtures in addition to the determination of parity.

The earliest development of the theory (Ha 48, Zi 50) gave a somewhat different but entirely equivalent form for the direction-polarization correlation function, which has been used extensively in experimental work. If the direction-direction correlation is expressed as in (I-6), then the second form of the direction-polarization correlation which we treat is obtained simply by replacing the coefficient $A_\nu(L_2L_2')$ in each term by§

$$A_\nu(L_2L_2')+(\pm)_{L_2'}a_\nu(L_2L_2')\cos 2\gamma. \quad \text{(I-11)}$$

The angle γ and the sign convention are the same as used in the foregoing. In this prescription the new coefficients a_ν can be expressed in terms of the old ones A_ν. Writing the series in $\cos^\nu\theta$ as well as in $P_\nu(\cos\theta)$, we have

$$W(\theta)=\sum A_\nu(L_2L_2')P_\nu(\cos\theta)=\sum Q_\nu(L_2L_2')\cos^\nu\theta \quad \text{(I-12)}$$

$$P_1W(\theta)=\sum(\pm)_{L_2'}a_\nu(L_2L_2')P_\nu(\cos\theta)$$
$$=\sum(\pm)_{L_2'}q_\nu(L_2L_2')\cos^\nu\theta \quad \text{(I-13)}$$

$$a_\nu(L_2L_2')=\sum_n \alpha_nA_n(L_2L_2') \quad \text{(I-14)}$$

$$q_\nu(L_2L_2')=\sum_n \beta_nQ_n(L_2L_2'). \quad \text{(I-14')}$$

The coefficients α_n and β_n depend on L_2, L_2', and ν. Numerical values for these coefficients are provided in Table III for pure multipoles, in Table IV for mixed radiation of the same parity, and in Table V for mixed radiation of opposite parity. We note here that $\alpha_0=\beta_0=\alpha_1=\beta_1=0$. Many examples of the use of this formula-

TABLE III(b). Numerical values of the coefficients in the series $q_\nu(LL)=\Sigma_n\beta_nQ_n(LL)$ (cf. Zi 50, Ha 48). For odd ν, all $\beta_n=0$.

L	ν	β_2	β_4	β_6
1	0	−1		
	2	1		
2	0	1	1	
	2	−1	−2	
	4	0	1	
3	0	2/3	0	−1/3
	2	−2/3	4	19/3
	4	0	−4	−7
	6	0	0	1

‡ Most authors give this condition in more elegant fashion by defining a quantity σ equal to 0 or 1 depending on the character of the radiation. Since there is no unanimity in the definition of this quantity, we prefer the foregoing statement (De 57a) in the hope of reducing confusion. The sign convention depends on the direction chosen for $\gamma=0$ and on the sign used in (I-7).

§ This substitution is obtained from the one used in the foregoing by expanding the associated Legendre function in terms of the ordinary functions.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

tion occur in the literature. We cite here the case of the pure octupole gamma ray in the reaction $F^{19}(p,\alpha)O^{16*}$, $(\gamma)O^{16}$, $E_\gamma=6.1$ Mev. At $E_p=0.33$ Mev it is known that the proton is captured in an s-state, so that the α-γ emission can be treated as an independent two-stage process. From the measurement of the direction-direction correlation (Ar 50, Ba 50) one selects the theoretical result

$$W(\theta)=1+111\cos^2\theta-305\cos^4\theta+225\cos^6\theta. \quad \text{(I-15)}$$

With the aid of Table III(b), one finds directly for the direction-polarization correlation for an octupole radiation

$$P_1W(\theta)=\pm\tfrac{1}{3}[(2Q_2-Q_6)-(2Q_2-12Q_4-19Q_6)\cos^2\theta -(12Q_4+21Q_6)\cos^4\theta+3Q_6\cos^6\theta]. \quad \text{(I-16)}$$

Inserting the numerical values of the direction-direction coefficients, one has

$$P_1W(\theta)=\pm[-1+131\cos^2\theta -355\cos^4\theta+225\cos^6\theta]. \quad \text{(I-17)}$$

The linear polarization is complete at $\theta=118°$ (or 62°), i.e., at that angle $J_0/J_{90}=0$ or ∞ depending on whether the radiation is magnetic or electric. The measurement of this quantity is discussed in Part III.

A problem frequently encountered is that of the dipole-quadrupole mixture in the radiation from a discrete level. From Tables III(a) and IV(a) we obtain

$$a_0(11)=-A_2(11) \quad a_0(12)=-\tfrac{1}{3}A_2(12) \quad a_0(22)=A_2(22)-\tfrac{1}{6}A_4(22)$$
$$a_2(11)=A_2(11) \quad a_2(12)=\tfrac{1}{3}A_2(12) \quad a_2(22)=-A_2(22)-\tfrac{5}{6}A_4(22)$$
$$a_4(22)=A_4(22). \quad \text{(I-18)}$$

TABLE IV(a). Numerical values of the coefficients in the series $a_\nu(LL')=\Sigma_n\alpha_nA_n(LL')$ for even ν.

L	L'	ν	α_2	α_4
1	2	0	−1/3	
		2	1/3	
1	3	0	−1/6	−1/6
		2	1/6	−5/6
		4	0	1
2	3	0	−1/2	−1/30
		2	1/2	−1/6
		4	0	1/5

TABLE IV(b). Numerical values of the coefficients in the series $q_\nu(LL')=\Sigma_n\beta_nQ_n(LL')$ for even ν (cf. Zi 50).

L	L'	ν	β_2	β_4
1	2	0	−1/3	
		2	1/3	
1	3	0	−1/6	
		2	1/6	−1
		4	0	1
2	3	0	−1/2	−2/5
		2	1/2	1/5
		4	0	1/5

TABLE V(a). Numerical values of the coefficients in the series $a_\nu(LL')=\Sigma_n\alpha_nA_n(LL')$ for odd ν.

L	L'	ν	α_3	α_5
1	2	1	−1	
		3	1	
1	3	1	−1/2	
		3	1/2	
2	3	1	3/2	−3/10
		3	−3/2	−7/10
		5	0	1

TABLE V(b). Numerical values of the coefficients in the series $q_\nu(LL')=\Sigma_n\beta_nQ_n(LL')$ for odd ν.

L	L'	ν	α_3	α_5
1	2	1	−1	
		3	1	
1	3	1	−1/2	
		3	1/2	
2	3	1	3/2	2
		3	−3/2	−3
		5	0	1

Inserting these quantities into (I-13), we find (changing the sign of the dipole term)

$$P_1W(\theta)=\pm[(A_2-\tfrac{4}{3}A_2(12)-\tfrac{1}{6}A_4) -(A_2-\tfrac{4}{3}A_2(12)+\tfrac{5}{6}A_4)P_2+A_4P_4]. \quad \text{(I-19)}$$

In this expression A_2 and A_4 are coefficients of the total correlation function, while $A_2(12)$ is the mixed coefficient. From this result we obtain at $\theta=90°$ and for even parity ($M1+E2$)

$$J_0=A_0+A_2+A_4-2A_2(12) \quad \text{(I-20)}$$
$$J_{90}=A_0-2A_2-\tfrac{1}{4}A_4+2A_2(12). \quad \text{(I-21)}$$

This example is a generalization of the one given above, (I-10″).

The third form of the direction-polarization correlation arises because of the convenience of discussing certain problems in terms of the magnetic sublevels of the radiating state as has been recognized by many authors. Because of the obvious physical insight provided by the method in the study of nuclear orientation, it has been the starting point of theoretical developments in this field (St 52a, To 53). Hence, we write

$$W(\theta)=\sum_m w(m)I_m{}^{LL}(\theta), \quad \text{(I-22)}$$

where $I_m{}^{LL}(\theta)$ is the intensity distribution of 2^L-pole radiation from the mth sublevel of the radiating state of angular momentum j, $w(m)$ is the relative population of this sublevel, and the summation over m is from $-j$ to j. If the angular momentum j_f of the final state is greater than zero, then we have further:

$$I_m{}^{LL}(\theta)=\sum_M |C(jLj_f;mM)|^2F_M{}^{LL}(\theta), \quad \text{(I-23)}$$

where the coefficient is the vector addition coefficient

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

TABLE VI(a). The functions $F_M^{LL'}(\theta)$ and $f_M^{LL'}(\theta)$, both given in the form $K\Sigma C_n P_n(\cos\theta)$. In each case $F_M^{LL'}(\theta)$ is tabulated on the first line and $f_M^{LL'}(\theta)$ on the second line. The upper sign corresponds to $M>0$; the lower to $M<0$.

L	L'	$\|M\|$	K	C_0	C_2	C_4	C_6
1	1	1	1	2	1		
				−1	1		
		0	1	2	−2		
				2	−2		
1	2	1	$\pm\frac{1}{3}\sqrt{15}$	0	3		
				−1	1		
		0	⋯	0	0		
				0	0		
2	2	2	1/21	42	−30	−12	
				−28	40	−12	
		1	1/21	42	15	48	
				7	−55	48	
		0	1/21	42	30	−72	
				42	30	−72	
2	3	2	$\pm(350)^{\frac{1}{2}}/210$	0	−30	30	
				14	−20	6	
		1	$\pm(35)^{\frac{1}{2}}/210$	0	60	150	
				−35	5	30	
		0	⋯	0	0	0	
				0	0	0	
3	3	3	1/66	132	−165	18	15
				−99	165	−81	15
		2	1/66	132	0	−42	−90
				−22	−110	222	−90
		1	1/66	132	99	6	225
				55	−121	−159	225
		0	1/66	132	132	36	−300
				132	132	36	−300

and $F_M^{LL}(\theta)$ is the intensity distribution of an (L,M)-pole. Associated with each $F_M^{LL}(\theta)$ there is of course a distribution function $f_M^{LL}(\theta)$ for linear polarization. Then,

$$F_M^{LL}(\theta,\gamma)=F_M^{LL}(\theta)\pm\cos 2\gamma f_M^{LL}(\theta) \qquad \text{(I-24)}$$

can be inserted in place of $F_M^{LL}(\theta)$ in (I-23) to produce the direction-polarization distribution (generalized to mixed radiation), namely,

$$W(\theta,\gamma)=\sum_m w(m)\sum_M\sum_{LL'}C(L)C(L')\delta_L\delta_{L'}$$
$$\times[F_M^{LL'}(\theta)\pm\cos 2\gamma f_M^{LL'}(\theta)]. \qquad \text{(I-25)}$$

(Agreement with previous notation is obtained if $\delta_L=1$). The $F_M^{LL'}(\theta)$ and $f_M^{LL'}(\theta)$ are given in Table VI.

There are several empirical examples of angular distributions which correspond to a single (L,M) pole. The reaction $F^{19}(p,\alpha)O^{16*}(\gamma)O^{16}$ has already been cited. In Table IV we see that the distributions given in (I-15, I-17) correspond to a pure (3,1)-pole. A second example is the capture reaction, $H^2(p,\gamma)He^3$. The measured directional distribution is fitted very closely by the function $F_0^{11}(\theta)\sim\sin^2\theta$. The polarization distribution is, therefore, [Table IV(b)]

$$F_0^{11}(\theta,\gamma)=(1\pm\cos 2\gamma)(1-\cos^2\theta). \qquad \text{(I-26)}$$

At $\theta=90°$ the polarization ratio $J_0/J_{90}=0$ or ∞ depending on whether the radiation is magnetic or electric. The measurement of this quantity is discussed in Part III.

For an angular distribution or correlation it is clearly possible to obtain the population numbers $w(m)$ from a knowledge of the preceding transitions (Li 49). In a nuclear orientation problem the population numbers are functions of the temperature and of the method used to produce orientation. In a given method they can also depend on specific parameters which may vary from one experiment to the next. The calculation of $w(m)$ has been treated very thoroughly in the literature and, insofar as possible, general formulas have been provided. In Table VII we list the methods of producing orientation which have been generally discussed and in the last column we give references which deal with the computation of $w(m)$.

For purposes of computation it is advantageous to retain as much as possible of the formulation which makes use of the functions $F_\nu(LL'j_nj)$. In an equation such as (I-1′) it is only necessary to replace the coefficient $A_\nu(1)$ by an orientation parameter $B_\nu(j)$ which describes the orientation of the initial state. Equation (I-1′) retains its generality. The orientation parameter is merely a linear combination of the population numbers (Gr 55a, Bl 57),

$$B_\nu(j)=(2\nu+1)^{\frac{1}{2}}\sum_m C(j\nu j;m0)w(m). \qquad \text{(I-27)}$$

For the linear polarization distribution we need only

TABLE VI(b). The functions $F_M^{LL'}(\theta)$ and $f_M^{LL'}(\theta)$, both given in the form $k\Sigma b_n\cos^n\theta$. In each case $F_M^{LL'}(\theta)$ is tabulated on the first line and $f_M^{LL'}(\theta)$ on the second line (cf. Ar 50, Zi 50). The upper sign corresponds to $M>0$; the lower to $M<0$.

L	L'	$\|M\|$	k	b_0	b_2	b_4	b_6
1	1	1	3/2	1	1		
				−1	1		
		0	3/2	2	−2		
				2	−2		
1	2	1	$\pm\frac{1}{2}\sqrt{15}$	−1	3		
				−1	1		
		0	⋯	0	0		
				0	0		
2	2	2	5/2	1	0	−1	
				−1	+2	−1	
		1	5/2	1	−3	4	
				1	−5	4	
		0	5/2	0	6	−6	
				0	6	−6	
2	3	2	$\pm(175/32)^{\frac{1}{2}}$	1	−6	5	
				1	−2	1	
		1	$\pm\frac{1}{8}\sqrt{35}$	1	−18	25	
				−1	−4	5	
		0	⋯	0	0	0	
				0	0	0	
3	3	3	7/32	15	−15	−15	15
				−15	45	−45	15
		2	7/32	10	−30	110	−90
				10	−110	190	−90
		1	7/32	1	111	−305	225
				−1	131	−355	225
		0	7/32	12	−132	420	−300
				12	−132	420	−300

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

the even quantities and of these the following will generally suffice:

$$B_0=1$$

$$B_2=[(1/5)j(j+1)(2j-1)(2j+3)]^{-\frac{1}{2}} \times[\sum_m m^2w(m)-(1/3)j(j+1)]. \quad \text{(I-28)}$$

The calculation of these quantities for the various experimental techniques is discussed in the references given in Table VII.

The orientation parameters are given in To 52 in a form which differs from the B's only in normalization. The expression for $\nu=2$ is

$$f_2=j^{-2}[\sum_m m^2w(m)-(1/3)j(j+1)]. \quad \text{(I-29)}$$

When the directional distribution has been obtained by any one of the foregoing techniques, the polarization distribution can be derived from it in the usual fashion with the substitution indicated by (I-4) or (I-11).

The polarization-polarization correlation has been discussed by several authors (Fa 48, Ma 51, Ll 52, Bi 53) and a general expression for it can be found in Ll 52. This type of correlation has not received much attention experimentally, since it provides very little additional information (Ha 48, Bi 53) to compensate for the increased experimental difficulties.

In a measurement of the γ-γ direction-polarization correlation, it has not always been feasible to distinguish the gamma rays in the detectors. It is then necessary to mix the $\gamma_1-\gamma_2$ direction-polarization function with the $\gamma_2-\gamma_1$ function according to the respective efficiencies of detection.

C. Circular Polarization Distribution

The γ-γ circular-circular correlation can be written in concise form (Go 46, Al 57) as

$$W(\theta p_1p_2)=\sum_\nu(-p_1)^\nu A_\nu(1)(-p_2)^\nu A_\nu(2)P_\nu(\cos\theta). \quad \text{(I-30)}$$

The summation is over the odd and even integers (including zero), and $p_n=\pm1$ according as the circular polarization of the nth radiation is right or left handed.‖ The coefficient $A_\nu(n)$ can be obtained from (I-2) which is valid for odd ν. The coefficients $F_\nu(L_nL_n'j_nj)$ for odd ν have been tabulated in Al 57.

If the expression (I-30) is summed over a p_n, the total intensity of the corresponding gamma ray is obtained. Thus, we find,

$$W(\theta)=\sum_{p_1p_2}W(\theta p_1p_2)=2\sum_{p_1}W(\theta p_1p_2)$$
$$=2\sum_{p_2}W(\theta p_1p_2)=\sum_{\nu\text{ even}}A_\nu(1)A_\nu(2)P_\nu(\cos\theta), \quad \text{(I-31)}$$

which is just the direction-direction correlation. Hence there are no direction-circular (or linear-circular) correlations. If, on the other hand, the excess of circular polarization (right handed minus left handed) is measured for either gamma ray or both, one obtains

$$P_3W(\theta)=\sum_{p_1p_2}p_1p_2W(\theta p_1p_2)=2\sum_{p_1}p_1p_2W(\theta p_1p_2)$$
$$=2\sum_{p_2}p_1p_2W(\theta p_1p_2)$$
$$=\sum_{\nu\text{ odd}}A_\nu(1)A_\nu(2)P_\nu(\cos\theta). \quad \text{(I-32)}$$

The quantity P_3 is the degree of circular polarization, as defined by Fano (Fa 49). It is analogous to the quantity P_1 defined for the linear polarization [see (I-5)]. In addition, it follows that

$$\sum_{p_1p_2}p_1W(\theta p_1p_2)=\sum_{p_1p_2}p_2W(\theta p_1p_2)=0,$$

which shows again there is no direction-circular correlation.

In the case of nuclear polarization produced by orientation at low temperatures, we merely identify the coefficient $A_\nu(1)$ in (I-32) with the orientation parameter $B_\nu(j)$. In terms of the population numbers $w(m)$, the most important parameter is (To 52, Bl 57)

$$B_1=2\sqrt{3}\sum_m mw(m).$$

or

$$f_1=j^{-1}\sum_m mw(m). \quad \text{(I-33)}$$

References which deal with the computation of these parameters are given in Table VII. The most important feature of B_1 is that its sign depends on the sign of the hfs interaction which produces the nuclear polarization. In turn the sign of the interaction depends on the sign of the nuclear magnetic moment. A measurement of the degree of circular polarization gives the sign in (I-32) and thus yields the sign of the magnetic moment.

TABLE VII. Some methods of producing nuclear orientation at low temperatures.

Method	Type of orientation	Applied field	References
External field (Brute force)	pol	$>10^5$ gauss	Go 34, Ku 35, Si 51, St 53a, Po 54, Kh 55, St 57a, Bl 57
Magnetic hfs (Rose-Gorter)	pol	$\sim10^2$ gauss	Go 48, Ro 49, Si 51, St 53a, Po 54, Gr 55a, Va 56, St 57a, Bl 57
Electric hfs (Pound)	align	none	Po 49, Si 51, St 53a, St 57a, Bl 57
Magnetic hfs (Bleaney)	align	none	Bl 51a, Si 51, St 53a, Po 54, Gr 55a, St 57a, Bl 57

‖ The term "right handed" is used for a photon with its spin in the direction of its momentum. This definition accords with recent practice, but (unfortunately) not with the "optical" definition.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

TABLE VIII. The functions $g_M{}^{LL'}(\theta)=K\Sigma C_n P_n(\cos\theta)$ which give the intensity distribution of circularly polarized radiation for an (LL',M) pole. The + or − sign corresponds to $M>0$ or $M<0$, respectively (cf. St 53).

L	L'	$\lvert M\rvert$	K	C_1	C_3	C_5
1	1	1	± 3	1		
		0	⋯	0		
1	2	1	$(3/5)^{\frac{1}{2}}$	3	2	
		0	$3/(5)^{\frac{1}{2}}$	2	−2	
2	2	2	± 1	2	−2	
		1	± 1	1	4	
		0	⋯	0	0	
2	3	2	$(14)^{\frac{1}{2}}/21$	−12	7	5
		1	$(35)^{\frac{1}{2}}/105$	48	7	50
		0	$(280)^{\frac{1}{2}}/105$	−18	−7	25
3	3	3	$\pm 1/6$	9	−14	5
		2	$\pm 1/6$	6	14	−20
		1	$\pm 1/6$	3	14	25
		0	⋯	0	0	0

The degree of circular polarization can be written in terms of the population numbers directly. In analogy to (I-25) for the case of linear polarization, we have

$$P_3W(\theta)=\sum_m w(m)\sum_M\sum_{LL'}C(L)C(L')\delta_L\delta_{L'}g_M{}^{LL'}(\theta), \quad \text{(I-34)}$$

where $C(L)$ stands for $C(jLj_f;mM)$. The function $g_M{}^{LL'}(\theta)$ gives the degree of circular polarization for an (LL',M) pole. A few of these functions are listed in Table VIII.

Circularly polarized radiation may be produced following the capture or emission of a polarized particle such as a neutron or a proton (Ha 51, Bi 51). The only case of importance so far has been the capture of polarized thermal neutrons as discussed in Part III. Since the orbital angular momentum is zero, there is no complication in applying the formalism of this discussion. In this case $A_1(1)$ in (I-32) becomes $\frac{1}{2}F_1(\frac{1}{2}\frac{1}{2}j_1j)$, by analogy with an absorbed gamma ray. The functions $F_1(\frac{1}{2}\frac{1}{2}j_1j)$ are tabulated in Table IX. As an example, consider $j_1=0$, $j=\frac{1}{2}$, $j_2=\frac{1}{2}$, and $L=1$. From Table IX we obtain $F_1(\frac{1}{2}\frac{1}{2}0\frac{1}{2})=-2.00$ and in Al 57 we find $F_1(11\frac{1}{2}\frac{1}{2})=-1.00$. Thus (I-32) becomes

$$P_3=+\cos\theta.$$

At $\theta=0$ (along the direction of the neutron polarization), $P_3=+1$ and the gamma rays are completely circularly polarized (right handed).

TABLE IX. Numerical values of $F_1(\frac{1}{2}\frac{1}{2}j_1j)$.

j_1	j	$F_1(\frac{1}{2}\frac{1}{2}j_1j)$	j_1	j	$F_1(\frac{1}{2}\frac{1}{2}j_1j)$	j_1	j	$F_1(\frac{1}{2}\frac{1}{2}j_1j)$
0	1/2	−2.000	2	3/2	0.894	7/2	3	1.000
1/2	1	−1.633	2	5/2	−1.366	7/2	4	−1.291
1	1/2	0.667	5/2	2	0.943	4	7/2	1.018
1	3/2	−1.491	5/2	3	−1.333	4	9/2	−1.277
3/2	1	0.816	3	5/2	0.976	9/2	4	1.033
3/2	2	−1.414	3	7/2	−1.309	9/2	5	−1.265

For the β-γ direction-circular correlation (Le 56, Al 57, Ga 57, Mo 57) the formal procedure is the same. For the beta particle, we write (Al 57)

$$A_\nu(\beta)=\sum_{L_1L_1'}F_\nu(L_1L_1'j_1j)b_\nu(L_1L_1'), \quad \text{(I-35)}$$

where $b_\nu(L_1L_1')$ is the particle factor for the beta ray. The mixing of multipole orders in the beta radiation is governed by the particle factors which are characteristic of the beta process. The particle factors for allowed beta transitions are given in Table X which is an extension of the table given by Alder *et al.* (Al 57). For the degree of circular polarization, we obtain from (I-32)

$$P_3=-[A_1(\beta)/A_0(\beta)]A_1(\gamma)P_1(\cos\theta),$$

where

$$A_1(\beta)=[F_1(11j_1j)\lvert M_{GT}\rvert^2S_1{}^{(11)}+F_1(01j_1j)\lvert M_{GT}\rvert\cdot\lvert M_F\rvert S_1{}^{(01)}],$$

$$A_0(\beta)=\lvert M_F\rvert^2S_0{}^{(00)}+\lvert M_{GT}\rvert^2S_0{}^{(11)}.$$

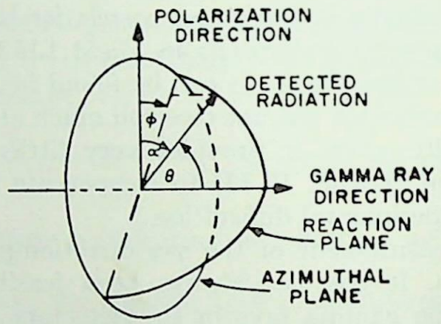


FIG. 1. Angles and vectors involved in a reaction sensitive to the linear polarization of a gamma ray, presented so as to apply to all such reactions discussed. The phrase "detected radiation" can represent a Compton scattered gamma ray, a photoproton from the photodisintegration of the deuteron, etc.

Let

$$S_0{}^{(00)}=\tfrac{1}{4}\lvert C_F\rvert^2,\quad S_0{}^{(11)}=\tfrac{1}{4}\lvert C_{GT}\rvert^2,\quad a=\lvert C_{GT}/C_F\rvert^2,$$

and $x=a^{\frac{1}{2}}M_{GT}/M_F$.

Then,

$$A_1(\beta)/A_0(\beta)=\tfrac{2}{3}(p/E)\times[F_1(11j_1j)Gx^2+F_1(01j_1j)Ia^{\frac{1}{2}}x]/(1+x^2),$$

where $\frac{2}{3}(p/E)G=S_1{}^{(11)}/S_0{}^{(11)}$ and $\frac{2}{3}(p/E)I=S_1{}^{(01)}/S_0{}^{(11)}$. This is the form and notation given in Bo 58. The experimental results on the beta-gamma correlation and on the electron distribution from polarized nuclei (Wu 57) have shown that $G=\pm 1$ for positrons and electrons. As an example, consider the Co^{60} decay, $5(\beta)4(2)2(2)0$. The beta and gamma radiations are pure. For the first gamma ray, with $G=-1$, we find

$$P_3=\tfrac{2}{3}(v/c)(0.774)(-0.645)\cos\theta=-\tfrac{1}{3}(v/c)\cos\theta$$

from the tables in Al 57. For the second gamma ray a factor must be inserted for the unobserved gamma ray $[\cos(J_1J_1')]$ but the numerical result is the same.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by eGangotri Foundation USA

TABLE X. Particle parameters $b_\nu(LL')$ for allowed beta transitions. The parameter $b_\nu(LL')$ is the product of the matrix element (ME) and $S_\nu^{(LL')}$. The coupling constants are denoted by C_S, C_V, etc. The energy and momentum of the electron are denoted by E and p, and α is the fine-structure constant. The Fierz interference terms have been put equal to zero (cf. Al 57, Mo 57, Ga 57).

L	L'	ME	$S_\nu^{(LL')}$
0	0	$\lvert M_F\rvert^2$	$S_0^{(00)}=\frac{1}{4}\{\lvert C_S\rvert^2+\lvert C_S'\rvert^2+\lvert C_V\rvert^2+\lvert C_V'\rvert^2\}$
1	1	$\lvert M_{GT}\rvert^2$	$S_0^{(11)}=\frac{1}{4}\{\lvert C_T\rvert^2+\lvert C_T'\rvert^2+\lvert C_A\rvert^2+\lvert C_A'\rvert^2\}$
			$S_1^{(11)}=\frac{1}{6}(p/E)\{2\,\mathrm{Re}(C_TC_T'^*-C_AC_A'^*)+(Z\alpha/p)\,\mathrm{Im}(C_TC_A'^*+C_T'C_A^*)\}$
0	1	$\lvert M_F\rvert\cdot\lvert M_{GT}\rvert$	$S_0^{(01)}=0$
			$S_1^{(01)}=\frac{1}{6}(p/E)\{\mathrm{Re}(C_SC_T'^*+C_S'C_T^*-C_VC_A'^*-C_V'C_A^*)+(Z\alpha/p)\,\mathrm{Im}(C_SC_A'^*+C_S'C_A^*-C_VC_T'^*-C_V'C_T^*)\}$

II. POLARIZATION-SENSITIVE PROCESSES

Primary emphasis is given in this part to those processes that have proved the most practicable and useful in the detection of polarization. They are presented in the approximate order of their importance as determined chiefly by the frequency of application. Section D, Part II, contains a brief discussion of some of the processes which have not as yet found use in the study of nuclear gamma rays.

A. Compton Effect

(a) *Sensitivity to Plane Polarization*

The sensitivity of the Compton effect to the polarization of an incident gamma ray can be obtained from the Klein-Nishina formula (Kl 29, Ni 29) which gives the differential cross section for the process. In almost all cases encountered experimentally the direction of polarization of the scattered photon is not of interest. Thus, for the present discussion the most useful form of the Klein-Nishina expression is the one in which a summation has been made over all directions of polarization of the scattered photon with the result¶

$$d\sigma_\phi=\frac{r_0^2}{2}\frac{k^2}{k_0^2}\left[\frac{k_0}{k}+\frac{k}{k_0}-2\sin^2\theta\cos^2\phi\right]d\Omega, \qquad \text{(II-1)}$$

where $r_0=e^2/m_0c^2$ is the classical radius of the electron, $d\Omega$ is the element of solid angle into which the photon is scattered, θ is the angle through which the incident photon is scattered, ϕ is the angle between the electric vector of the incident photon and the plane of scattering, and k_0 and k are the energies of the incident and scattered photons, respectively. These energies are related by the expression

$$k=\frac{k_0}{1+(k_0/m_0c^2)(1-\cos\theta)}. \qquad \text{(II-2)}$$

The angles and vectors involved in a Compton scattering are illustrated in Fig. 1.

The effect of polarization is contained in the $\cos^2\phi$ factor of the last term of (II-1) which shows that the scattering is a maximum in the plane normal to the electric vector of the incident photon. As the energy of the incident gamma ray approaches zero the sum of the first two terms in the brackets approaches two. Thus in the limit of zero energy at $\theta=90°$ the Compton effect has an ideal response to polarization, i.e., the differential cross section is zero when $\phi=0°$ and a maximum when $\phi=90°$. At any given θ, an increase in the energy k_0 produces a decrease in the efficiency of analysis as expressed by the asymmetry ratio, $R=d\sigma_{90}/d\sigma_0$. Of especial interest is $\theta_{\max}$, the value of θ which makes R a maximum for a given energy k_0. This $\theta_{\max}$ is 90° for $k_0=0$ and decreases with increasing energy as shown in Fig. 2, which gives values of $\theta_{\max}$ up to an energy of about 6 Mev. A similar curve for energies up to about 1.5 Mev is given in Me 50.**

Figure 2 also shows the values of R at $\theta=\theta_{\max}$ plotted

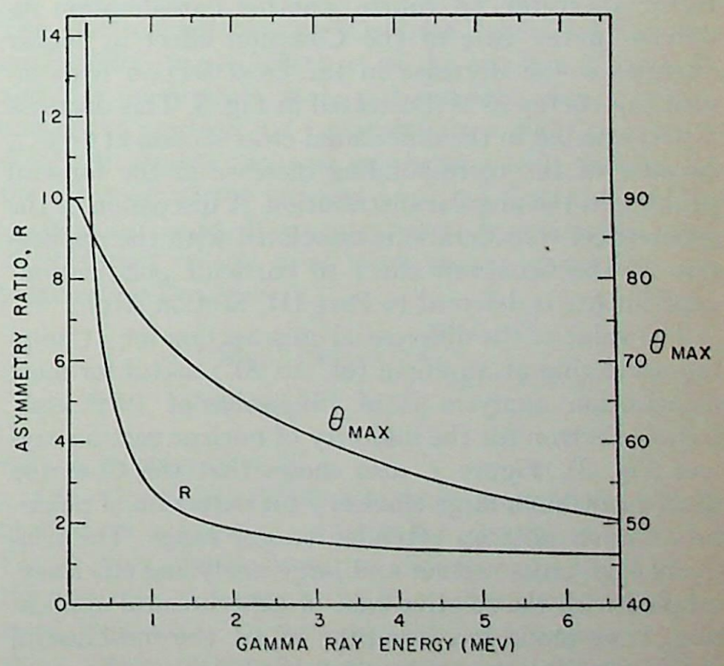


FIG. 2. The Compton scattering angle $\theta_{\max}$ at which the asymmetry ratio R is maximum, and the asymmetry ratio R for $\theta=\theta_{\max}$ as a function of incident photon energy up to about 6 Mev, the highest energy at which a Compton polarimeter has been used.

¶ A more detailed treatment including the derivation of this and other forms of the differential cross section can be found in He 44; see also Ev 55.

** There is disagreement between the latter curve and the one given here, which becomes larger as the energy increases from zero. The reason for this discrepancy is not known: perhaps an approximation valid only in the region of small energy was made in the calculations for the curve in Me 50. In the curve presented here no approximations have been made.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

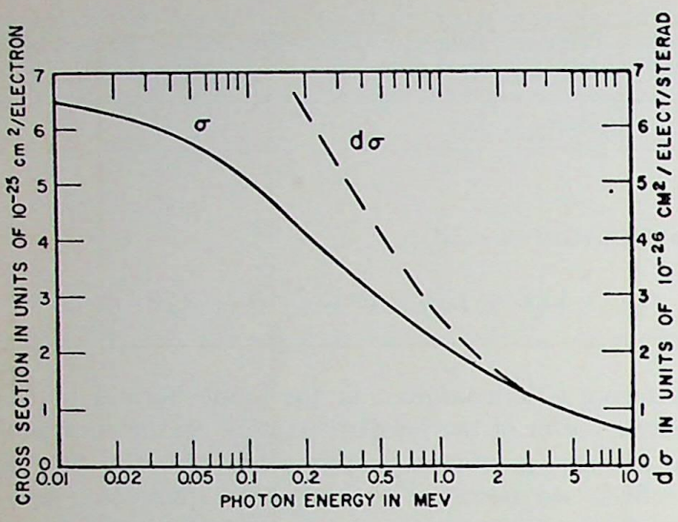


FIG. 3. Total cross section σ for the Compton effect (curve from Da 52a), and the differential cross section $d\sigma$ at $\theta=\theta_{max}$, as a function of incident photon energy.

as a function of energy. This figure shows that beyond about 2 Mev the value of R decreases very slowly as it approaches a value of one at infinite energy. Consequently, the Compton effect might in some cases be useful at energies as high as 10 Mev since at least a 10% effect would exist. However, its use at this energy might be somewhat marginal since the values of R given in Fig. 2 are for an ideal geometry. In any practicable experimental arrangement the value of R is considerably reduced by the angular spread of the detector apertures. Of course, another consideration involved in the case of the Compton effect at higher energies is the decrease in the cross section with increasing energy as is illustrated in Fig. 3. This decrease is accentuated in the differential cross section at $\theta=\theta_{max}$ because of the corresponding increase in the forward peaking in the angular distribution. A discussion of the geometrical considerations associated with the application of the Compton effect to particulr polarization experiments is deferred to Part III, Section A(c).

The value of the differential cross section for a Compton scattering at an angle (65° to 90°) useful for most polarization analyses is of the order of 10^{-26} cm^2/sterad/electron for the majority of nuclear gamma rays (see Fig. 3). Figure 2, also shows that the Compton effect furnishes a large efficiency for detection of polarization over quite an extended energy range. The relatively high cross section and large analyzing efficiency, coupled with the relative ease of experimental application, have made the Compton effect the most useful of the polarization-sensitive mechanisms.

(b) Sensitivity to Circular Polarization

Interest in the sensitivity of the Compton effect to circular polarization has increased considerably with the recent investigations of nonconservation of parity in weak interactions by means of experiments on the circular polarization of gamma rays following beta decay. Although the scattering of circularly polarized gamma rays by polarized electrons had been discussed by earlier workers (Bi 51, Ha 51, Cl 52), the first effective experimental study of the effect was made by Gunst and Page (Gu 53). In this work the total cross section is given as $\sigma=\sigma_0\pm\sigma_1$ where σ_0 and σ_1 are the parts unaffected and affected, respectively, by circularly polarized gamma rays. Gunst and Page measured the cross section for the effect at $k_0=2.62$ Mev and found a value of $\sigma_1/\pi r_0^2=0.089\pm0.007$ in good agreement with the theoretical value, 0.093, found from expressions calculated by Stehle (see Gu 53). The experiment was done by measuring the difference in intensity of the 2.62-Mev gamma rays from ThC″ which were transmitted through an iron bar 30-cm long when magnetized, and then demagnetized. Although the gamma rays were initially unpolarized, they can be considered as consisting of equal components of right and left circularly polarized radiation. Consequently, when the magnetic field is applied, one of these components is preferentially scattered.

Lipps and Tolhoek (Li 54, Li 54a; see also Fr 38, Fa 49, To 56), have derived the Compton cross section in a general form with all polarizations taken into account,

$$d\sigma=r_0^2(k^2/k_0^2)\Phi(\mathbf{k}_0,\mathbf{k},\xi^0,\xi,\zeta^0,\zeta)d\Omega, \qquad \text{(II-3)}$$

where ξ^0, ξ, ζ^0, and ζ are the respective polarization vectors of the incident and scattered photon and of the electron before and after interaction, and $\Phi(\mathbf{k}_0,\mathbf{k},\xi^0,\xi,\zeta^0,\zeta)$ is a linear function of these polarization vectors. The function Φ can be separated into 16

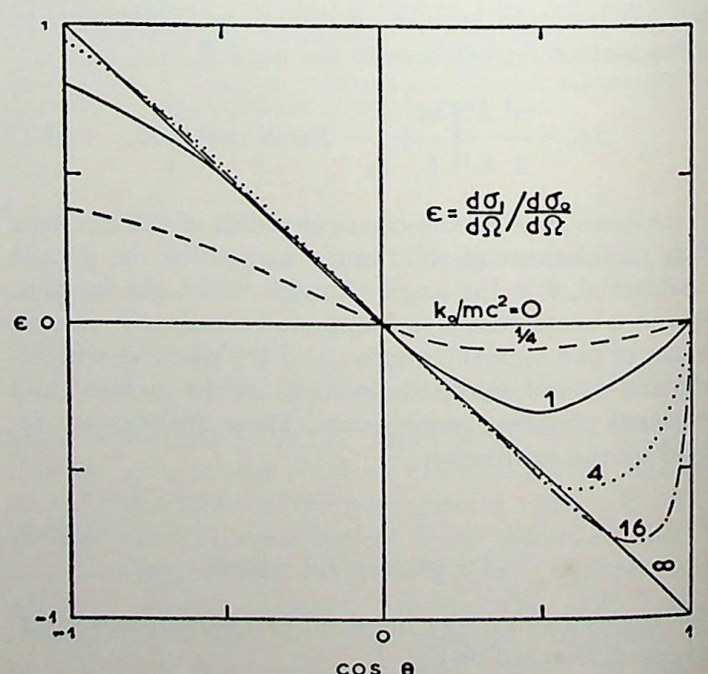


FIG. 4. Polarization-sensitive part $d\sigma_1$ of the differential Compton cross section for electrons polarized along the axis of incident circularly polarized photons. The curves are normalized to the polarization-insensitive cross section $d\sigma_0$ and plotted for several energies as a function of cosθ, where θ is the angle through which the photon is scattered (figure from Gu 53).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

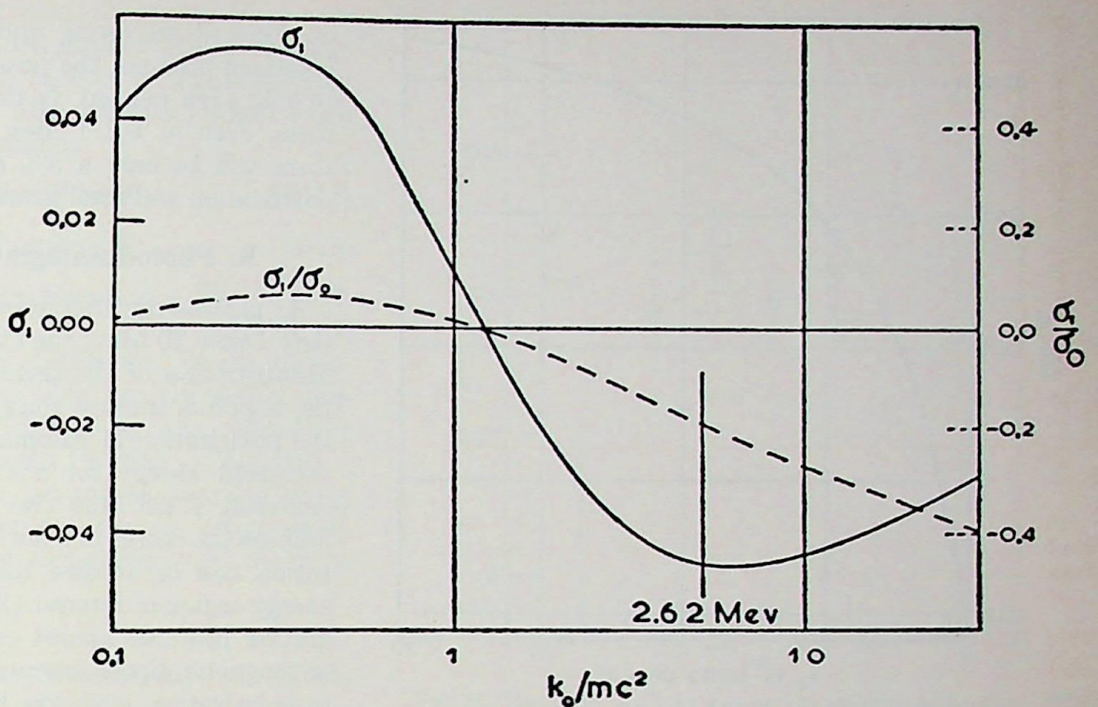


FIG. 5. Polarization-sensitive part σ_1 of the total Compton cross section (in units of $2\pi r_0^2$) for the same electron and photon polarizations as in Fig. 4. In the dashed curve σ_1 is normalized to σ_0, the polarization-insensitive part of the total cross section. The curves are plotted against incident photon energy (figure from Gu 53).

terms, depending on the 16 different ways of choosing sets of polarization vectors. Thus,

$$\Phi=\Phi_0+\Phi_1+\Phi_2+\Phi_3+\Phi_4, \qquad \text{(II-4)}$$

where Φ_0 is independent of polarizations,

$$\Phi_1=\Phi_1(\xi^0)+\Phi_1(\xi)+\Phi_1(\zeta^0)+\Phi_1(\zeta), \qquad \text{(II-5)}$$

$$\Phi_2=\Phi_2(\xi^0,\xi)+\Phi_2(\zeta^0,\zeta)+\Phi_2(\xi^0,\zeta^0)+\Phi_2(\xi,\zeta)+\Phi_2(\xi^0,\zeta)+\Phi_2(\xi,\zeta^0), \qquad \text{(II-6)}$$

and Φ_3, and Φ_4 depend on 3 and 4 polarization vectors, respectively. Explicit expressions are given in Li 54a for all the $\Phi_0,\cdots,\Phi_4$.

Thus far, the only experiments performed have involved not more than two polarizations. It is necessary then to average over unobserved initial polarizations and sum over unobserved final polarizations in order to obtain the cross section for some definite experiment. Since the states represented by $-\xi$ and $-\zeta$ are orthogonal to those represented by ξ and ζ, the terms which are functions only of the unobserved polarizations will vanish in these averages and sums.

For the case of interest here the initial photon polarization ξ^0 and the initial electron polarization ζ^0 are specified. For this case, (II-3) takes the form,

$$d\sigma(\xi^0,\zeta^0)=4r_0^2(k^2/k_0^2)\times[\Phi_0+\Phi_1(\xi^0)+\Phi_1(\zeta^0)+\Phi_2(\xi^0,\zeta^0)]d\Omega. \qquad \text{(II-7)}$$

In this expression Φ_0 gives the usual Klein-Nishina formula independent of polarization, $\Phi_1(\zeta^0)\equiv 0$, and $\Phi_1(\xi^0)=0$ if no linear polarization of the gamma ray is present. Hence, with

$$d\sigma=d\sigma_0+d\sigma_1 P \qquad \text{(II-8)}$$

we have (To 56)

$$Pd\sigma_1=4r_0^2(k^2/k_0^2)\Phi_2(\xi^0,\zeta^0)d\Omega=-\tfrac{1}{2}Pr_0^2\left(\frac{k^2}{k_0^2}\right)\left(\frac{k_0}{k}-\frac{k}{k_0}\right)\cos\theta d\Omega. \qquad \text{(II-9)}$$

The electron polarization has been taken parallel (or antiparallel) to the direction of the incident photon,†† and P is the product of the polarizations of the photon and of the electron. (P is positive for a right circularly polarized gamma ray and an electron spin parallel to the direction of the incident photon.) Equation (II-9) is given in Gu 53. If an integration is performed over $d\Omega$, the part of the total cross section dependent on polarization becomes

$$\frac{\sigma_1}{2\pi r_0^2}=\frac{1+4k_0^2+5k_0^2}{k_0(1+k_0)^2}-\frac{1+k_0}{2k_0^2}\ln(1+2k_0). \qquad \text{(II-10)}$$

with k_0 in units of m_0c^2.

The ratio of the differential cross sections $d\sigma_1/d\sigma_0$ is plotted as a function of $\cos\theta$ for several energies in Fig. 4. This graph shows the difference in the sign of the response to polarization in the forward and backward quadrants. In general, also, the response at a given angle increases with increasing energy. The polarization-dependent total cross section σ_1 is presented as a function of energy in Fig. 5. The change in sign of σ_1 at $E=1.25m_0c^2=0.65$ Mev arises as a result of the difference in sign of $d\sigma_1$ for forward and backward directions together with the fact that the forward Compton scattering becomes more dominant at higher energies.

†† This condition is easily removed (see Gu 53 and Li 54a).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

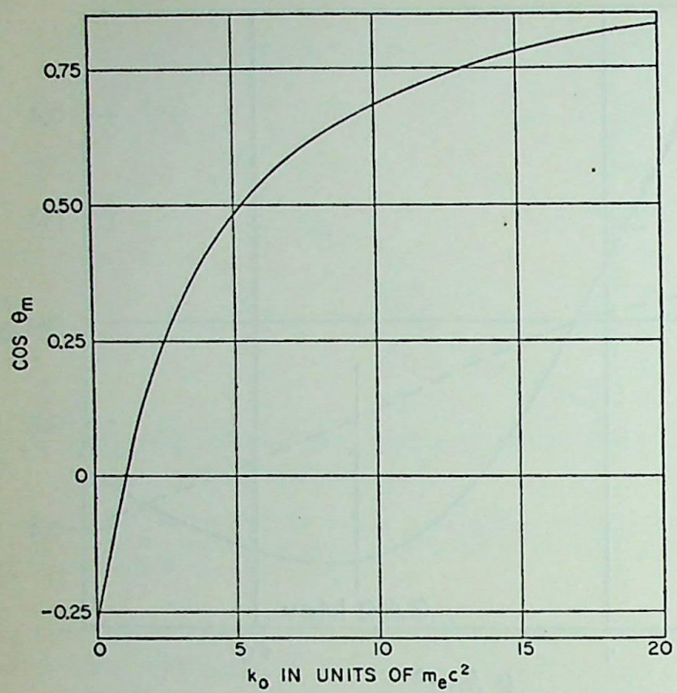


FIG. 6. Cosine of the Compton scattering angle which yields the maximum relative azimuthal anisotropy due to circular polarization, plotted as a function of the incident photon energy. In this case the circular polarization is perpendicular to the electron polarization (figure from Be 57).

This discussion has led to expressions applicable to the case in which circularly polarized gamma rays are scattered from electrons polarized parallel or antiparallel to the initial direction of propagation of the gamma ray. This is the case most used so far in experiments on circular polarization of gamma rays. Virtually all of these experiments detect the difference in the intensity of gamma rays transmitted through, or scattered forward, from magnetized iron, when the magnet is turned on and when it is turned off or reversed.

Recently Beard and Rose (Be 57) have suggested the use of the case in which the circularly polarized gamma rays are scattered from electrons polarized perpendicular to the initial direction of gamma-ray propagation. This case offers the advantage of determining an azimuthal asymmetry from two simultaneous measurements at opposite azimuthal scattering angles in the plane of the electron spin and photon propagation vector. Beard and Rose show‡‡ that at a certain scattering angle $\theta=\theta_m$, the relative azimuthal anisotropy J_1/J_0 is a maximum. Figure 6 gives a graph of $\cos\theta_m$ as a function of photon energy k_0. In Fig. 7, J_1/J_0 at $\theta=\theta_m$ is shown as a function of the energy of the incident photon. For completely polarized electrons, J_1/J_0 reaches a maximum value of 0.33 at 0.511 Mev.

The most significant practical limitation involved in applying the above cases to an experiment is the fact that only about 2 electrons of the 26 in an iron atom are polarized. Thus in all of the experiments performed on forward scattering and transmission of circularly polarized photons, the possible effects amount at most to only a few percent. In the case treated by Beard and Rose, even at 0.511 Mev where J_1/J_0 is maximum, there will be only a 5% effect (for complete photon polarization and ideal geometry).

‡‡ The cross section may be obtained in the manner that led to (II-9) except that the electron polarization is taken perpendicular to the direction of the incident photon.

B. Photodisintegration of the Deuteron

At gamma-ray energies below 20 Mev, and particularly below 10 Mev, the existing theory for the photodisintegration of the deuteron is satisfactory. This is the region of interest since we are concerned only with the polarization of gamma rays of nuclear origin. The threshold energy for the photodisintegration of the deuteron, 2.225 Mev (Be 56), places a definite lower limit on the energy of a nuclear gamma ray whose polarization can be studied using this mechanism. In the energy region of interest (2.225 to 20 Mev) the disintegration proceeds almost entirely by an electric dipole or magnetic dipole interaction. The former is sensitive to polarization, while the latter is not.

The cross section for the photodisintegration of the deuteron in the form derived by Bethe and Longmire (Be 50 and Be 56), assuming a central force potential, is expected to be valid below about 10 Mev. The differential cross section for electric dipole disintegration, the one that is polarization-sensitive is given by

$$d\sigma^e=2\left(\frac{e^2}{\hbar c}\right)\left[\frac{\gamma k^3}{(\gamma^2+k^2)^3}\right]\left(\frac{1}{1-\gamma r_{0t}}\right)\cos^2\alpha d\Omega, \quad \text{(II-11)}$$

where α is the angle between the direction of polarization of the gamma ray and the direction of motion of the proton (or neutron), and r_{0t} is the effective neutron-

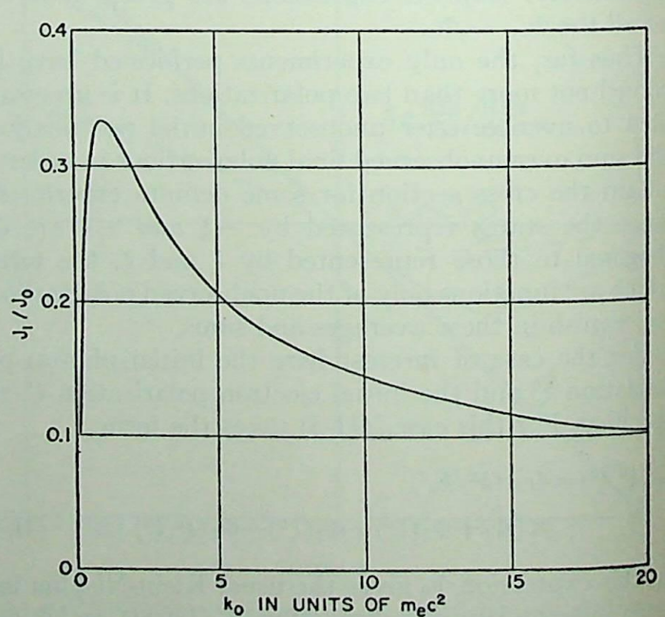


FIG. 7. Maximum relative azimuthal anisotropy J_1/J_0 resulting from Compton scattering of incident circularly polarized photons, plotted as a function of incident photon energy. The circular polarization is perpendicular to the electron polarization (figure from Be 57).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by eGangotri S3 Foundation USA

proton range in the triplet state. The parentheses containing r_{0t} represents a correction to the zero-range approximation for the cross section, which is given by the rest of the expression. The quantities γ and k can be obtained from the expressions

$$E=h\nu-W_1=\hbar^2k^2/M \qquad \text{(II-12)}$$

$$W_1=\hbar^2\gamma^2/M, \qquad \text{(II-13)}$$

where $h\nu$ is the energy of the gamma ray, W_1 the binding energy of the deuteron, and M the mass of the nucleon. That the angular distribution of the photoprotons varies as $\cos^2\alpha$ has been experimentally confirmed by Wilkinson (Wi 52) using the completely polarized gamma rays from the reaction $D_2(p,\gamma)He^3$.

Since it is usually the azimuthal distribution of photoprotons or photoneutrons that is observed in determining the polarization of a gamma ray, the more useful form of (II-11) is

$$d\sigma_{\phi^e}=\frac{2e^2}{hc}F(\gamma,k)\sin^2\theta\cos^2\phi d\Omega, \qquad \text{(II-14)}$$

where θ and ϕ are the usual polar coordinates. The direction of propagation of the gamma ray is given by $\theta=0$, and its direction of polarization by $\theta=90°$, $\phi=0$. The geometrical relationship of θ, ϕ, and α is illustrated in Fig. 1.

If the photodisintegration proceeded entirely by an electric dipole transition, the reaction would produce an ideal response to polarization in the sense defined above, namely that $R=d\sigma_{90}/d\sigma_0=0$.§§ Over a large part of the energy range of interest this is nearly the case. However, as the gamma-ray energy is decreased to the threshold value the reaction proceeds increasingly

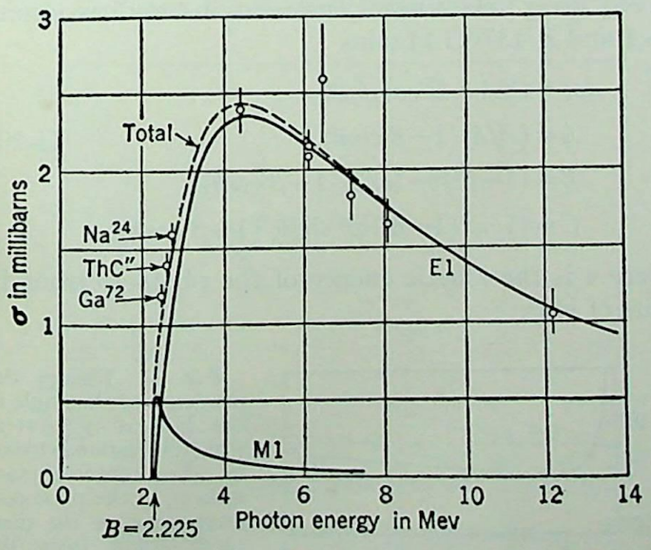


FIG. 8. Total cross section for the photodisintegration of the deuteron as a function of incident photon energy. Solid curves show the electric dipole and magnetic dipole cross sections as labeled, and dashed curve shows their sum. Points show representative experimental data (figure from Ev 55).

§§ In the Compton effect we have $R=\infty$ for an ideal response.

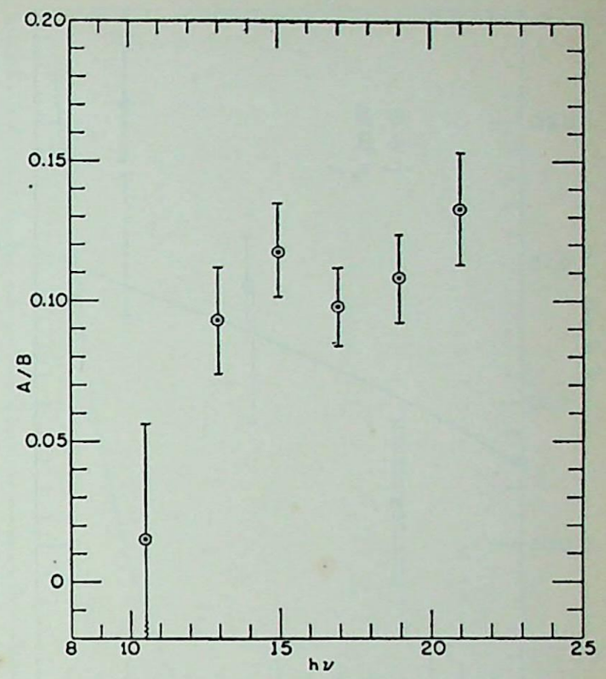


FIG. 9. The ratio A/B as a function of energy in Mev. This ratio is a measure of the isotropic term in the cross section for the photodisintegration of the deuteron. A noticeable effect starts at about 10 Mev (figure from Wh 58).

by the magnetic dipole interaction which yields an isotropic distribution of photoprotons. The total cross section for this process is given by (Se 53 and Be 50)

$$\sigma^m=\frac{2\pi}{3}\frac{e^2}{hc}\left(\frac{h}{Mc}\right)^2(\mu_n-\mu_p)^2\frac{k\gamma[\gamma-1/a_s]^2}{(k^2+\gamma^2)(k^2+1/a_s^2)}R(E), \qquad \text{(II-15)}$$

where μ_n and μ_p are the magnetic moments of the neutron and proton, respectively, a_s is the singlet scattering length, and $R(E)$ is a correction for the nonzero range of the neutron-proton force, which has been derived by Salpeter (Sa 51).

Since the presence of the magnetic dipole interaction diminishes the polarization sensitivity of the reaction as a whole, it is of interest to compare the total cross sections for the magnetic and electric processes as a function of the gamma-ray energy. The total cross section for the electric dipole process, obtained by integrating (II-11) over θ and ϕ, is

$$\sigma^e=\frac{8\pi}{3}\left(\frac{e^2}{\hbar c}\right)\left[\frac{\gamma k^3}{(\gamma^2+k^2)^3}\right]\left(\frac{1}{1-\gamma r_{0t}}\right). \qquad \text{(II-16)}$$

The two cross sections in the zero-range approximation (i.e., without the range correction given by the last factor in each expression) are compared in Fig. 8 (Ev 55). The sum of the two cross sections is also shown as are several experimental points which agree fairly well with the curve and indicate the extent of validity of the zero-range approximation.

Other effects, which generally tend to diminish the ideal polarization sensitivity of the electric dipole dis-

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

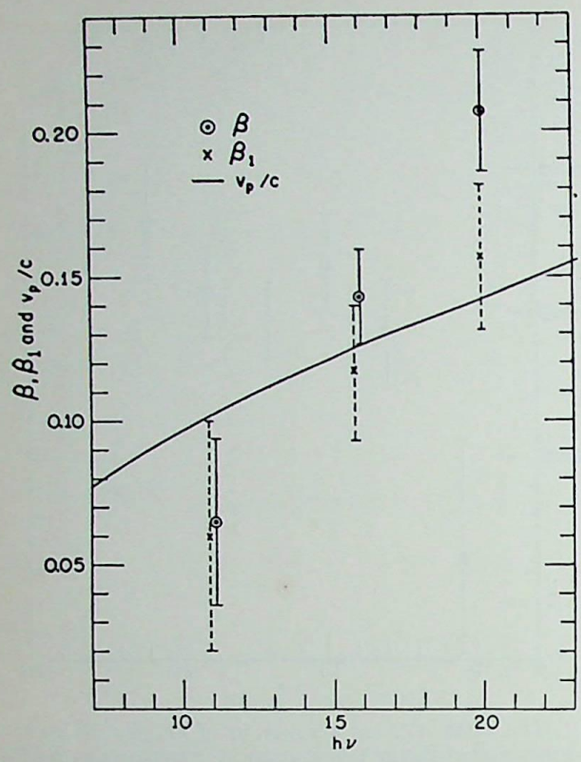


FIG. 10. The asymmetry coefficients β and β_1, as defined by (II-17) and (II-18), as a function of incident photon energy. Values of β and β_1 are obtained after first determining A/B. For comparison the theoretical prediction $v/c=[(h\nu-W_1)/Mc^2]^{\frac{1}{2}}$ (Ma 49) is plotted (figure from Wh 58).

integration, arise at energies above 10 Mev, and especially above 20 Mev. Since the nature of these effects is under current theoretical discussion, (Hu 57a) we confine attention to the empirical situation. The cross section is given in either one of the following forms (Wh 58):

$$d\sigma/d\Omega=A+B\sin^2\theta(1+2\beta\cos\theta), \qquad \text{(II-17)}$$

$$d\sigma/d\Omega=(A+B\sin^2\theta)(1+2\beta_1\cos\theta). \qquad \text{(II-18)}$$

The term in β or β_1 is attributed to the onset of the electric quadrupole interaction. In the recent experimental work by Whetstone and Halpern (Wh 58), values of A/B and β or β_1 have been determined as shown in Figs. 9 and 10. These values of A/B, along with those of Allen (Al 55), are in good agreement with the recent calculations of de Swart and Marshak (De 58).

All these expressions are given in the center-of-mass system. Thus, in addition to the forward asymmetry which appears at higher energies, a forward asymmetry arises experimentally in the laboratory system because of the forward momentum of the gamma ray. This effect is illustrated in Fig. 11 which gives the laboratory angle corresponding to 90° in the center-of-mass system. In the range of interest the shift is not large.

The magnetic dipole reaction and the deviations from the $\cos^2\alpha$ dependence above 10 Mev constitute the only significant effects that diminish the efficiency for detection of polarization in the energy region of interest here. In fact, from about 4 to 12 Mev the asymmetry ratio $R=d\sigma_{90}/d\sigma_0$ is less than 0.04. Perhaps the only serious drawback to the use of the photodisintegration of the deuteron in the detection of polarization is the relatively low total cross section ($\sim 10^{-27}$ cm^2; see Fig. 8).

C. Photoelectric Effect

The earliest measurements of polarization in the kev range made use of the photoelectric effect. Experiments using polarized photons having energies less than 40 kev (Wi 23, Bu 24, Ki 31) confirmed the basic theoretical prediction (Au 27, So 30) that the distribution of photoelectrons should vary approximately as $\cos^2\alpha$. As in the case of the photodisintegration of the deuteron, this distribution in terms of polar angles is $\sin^2\theta\cos^2\phi$, where α, θ, and ϕ are defined in the same way as before (Fig. 1).

The distribution has been calculated for electrons ejected from the K shell. Correspondingly, almost all the work dealing with the polarization sensitivity of the photoelectric effect has been concerned with emission from the K shell,|| || since it constitutes about 80% of the total emission for the energies of interest here (below about 1 Mev).

Fischer (Fi 31) and Heitler (He 44) have given the differential cross section in a nonrelativistic form (good to energies of about $\frac{1}{2}m_0c^2$):

$$d\sigma_\phi=r_0^2Z^5\frac{(32)^{\frac{1}{2}}}{137^4}\left(\frac{m_0c^2}{k_0}\right)^{7/2}\times\frac{\sin^2\theta\cos^2\phi}{[(1+k_0/2m_0c^2)^2-\beta\cos\theta]^4}d\Omega, \qquad \text{(II-19)}$$

where $\beta=v/c$ and Z is the atomic charge. Sauter (Sa 31, Sa 55) using relativistic considerations and assuming $\beta\sim 1$ and $Z/137\ll 1$ obtains

$$\begin{aligned} &d\sigma_\phi\sim C(A+B\cos^2\phi)d\Omega,\\ &A=(\epsilon^2/4)(1-\beta\cos\theta), \qquad \text{(II-20)}\\ &B=(1-\beta^2)^{\frac{1}{2}}-(\epsilon/2)(1-\beta\cos\theta)\\ &C=(1/\epsilon)^4(1-\beta^2)^{\frac{1}{2}}\beta^2\sin^2\theta/(1-\beta\cos\theta)^4, \end{aligned}$$

where ϵ is the kinetic energy of the photoelectrons in units of m_0c^2.

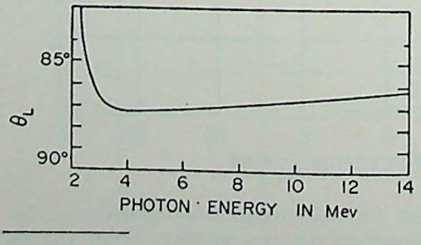


FIG. 11. Energy dependence of the angle in the laboratory system which corresponds to 90° in the center-of-mass system, in the photodisintegration of the deuteron (data from Wi 51a).

|| || A nonrelativistic calculation by Schur (Sc 30) gives an azimuthal distribution for the electrons ejected from the L_I level similar to the nonrelativistic K-shell distribution. However, the emissions from the L_I shell represent only a small fraction of those emissions (20%) which are not from the K shell.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

In these expressions there is a pronounced forward asymmetry in the angular distribution which increases as the energy of the gamma ray is increased. This effect is illustrated in Fig. 12 which gives theoretical curves for the photoelectron distribution as a function of θ for several energies. In the nonrelativistic case the response to polarization is ideal, i.e., $R=d\sigma_{90}/d\sigma_0=0$. In the relativistic case, however, the response is decreased by the fractional amount A/B. Another effect predicted by (II-20) is that the favored direction of emission of the photoelectron changes by 90° as the gamma-ray energy passes through 0.51 Mev. This is shown in Fig. 13.¶¶ Accordingly, above 0.51 Mev the photoelectron would be emitted preferentially normal to the direction of polarization of the gamma ray.

A series of measurements with polarized photons (>40 kev) designed to investigate this and other features of Sauter's expression has been carried out by Hereford and co-workers. Hereford (He 51, He 51a) studied the azimuthal distribution of the photoelectrons ejected from lead by the polarized radiation from positron-electron annihilation, and found a much greater asymmetry in the azimuthal distribution than predicted by (II-20). In a revised experiment Hereford and Keuper (He 53) obtained a result in rough agreement with that predicted by Sauter, if his equation, which is valid chiefly for low Z, is applied to the case of lead. McMaster and Hereford (Mc 54) continued the experiments by measuring the azimuthal asymmetry using photons (from Co^{60}) which had been polarized by Compton scattering. By changing the angle of scattering they were able to obtain polarized gamma rays of different energies. Their measured values of R fell below unity above about 0.55 Mev, showing fair agreement with the theoretical curve in Fig. 13.

Brini *et al.* (Br 57) recently repeated the experiment

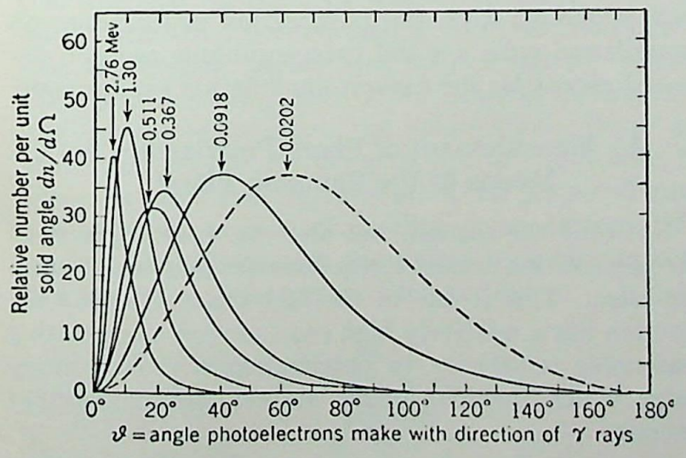


FIG. 12. Differential cross sections for photoelectrons ejected from the K shell, plotted against the angle of emission. The curves for the different photon energies are not normalized with respect to each other. Solid curves are calculated from Sauter's relativistic formula (Sa 31), the dashed curve from Fischer's non-relativistic formula (Fi 31) (figure from Ev 55).

¶¶ In this figure it is the reciprocal of R (as previously defined) that is plotted.

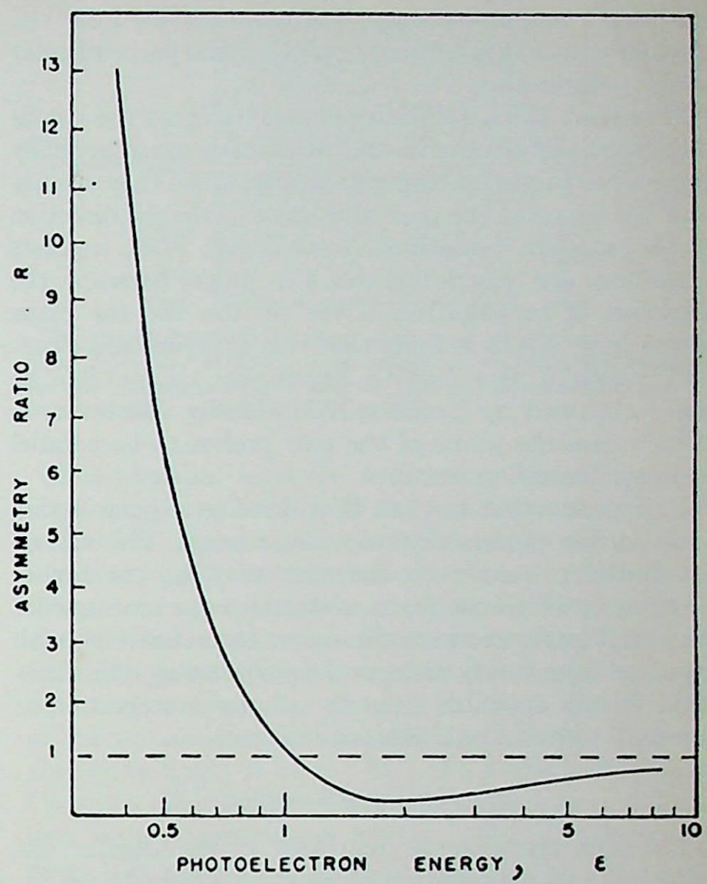


FIG. 13. Asymmetry ratio $R=d\sigma_0/d\sigma_{90}$ for photoelectrons ejected from the K shell by linearly polarized photons for $\theta=90°$, as calculated from Sauter's formula (Sa 31). The abscissa is the incident photon energy in units of m_0c^2 (figure from Mc 54).

of McMaster and Hereford, but did not find the 90° shift in the polarization response. Their data agree better with a theoretical curve of Archibald as given in Br 57. Apparently the knowledge of the polarization sensitivity of the photoelectric effect at higher energies is still somewhat uncertain. However, it is clear that the reaction can be used for gamma-ray energies which are well below 0.5 Mev.

The primary disadvantage of the photoelectric effect as an analyzer of polarization is that it requires detection of the photoelectron. Low-energy electrons are somewhat more difficult to detect than are the gamma rays in a typical Compton process. Furthermore, the scattering of low-energy electrons will present a serious problem.

D. Other Polarization-Sensitive Processes

(a) *Pair Production*

Berlin and Madansky (Be 50a) have investigated the possibility of using pair production as a polarization analyzer. Assuming a perfectly plane polarized incident photon they have calculated the expected asymmetry ratio $R=d\sigma_{90}/d\sigma_0$ for several different experimental cases. They prefer the experimental arrangement in which all pairs in a particular plane are counted re-

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

gardless of angle or energy. For this case $R=1.23$; i.e., the pair emission is a maximum in a plane perpendicular to the polarization.

However, Wick (Wi 51) pointed out that the conditions specified by Berlin and Madansky are practically impossible to realize experimentally, since they specify that the plane of the pair contains exactly the direction of the incident quantum. Accordingly, Wick suggests removing the restriction on the angle between the direction of propagation of the photon and the plane of the pair. He then finds that this experimental situation yields a polarization sensitivity comparable to those obtained by Berlin and Madansky. However, in Wick's case the plane of the pair prefers to be parallel to the polarization vector.

Pair production has not been used as a polarization analyzer for gamma rays of nuclear origin. This can be attributed primarily to the fact that, at the higher energies at which the reaction would prove most useful, the small angle between the pairs, the relatively small expected asymmetry ratio, and the scattering difficulties make it less desirable than the photodisintegration of the deuteron as a polarization analyzer.

(*b*) *Nuclear Photoeffect*

Although the general treatment of the angular distribution and polarization of reaction products (Bl 51, Si 53) is applicable to photonuclear reactions, Agodi (Ag 57) has given a specific treatment of the polarization sensitivity of photonuclear reactions. Without using a model for the nucleus for the photonuclear reaction, Agodi shows on quite general grounds (conservation of parity and angular momentum) that the azimuthal dependence of the differential cross section assumes a form proportional to $(1+\alpha\cos^2\phi)$, where ϕ is the azimuthal angle as defined in Fig. 1. This result is quite general regardless of whether a single or mixed multipole transition is involved in the photonuclear process. In the case in which a single multipole is reponsible for the transition, the value of α for a completely polarized incoming gamma ray can be determined from the angular distribution (in θ) coefficient, if it is known. Such is not generally the case, however, when two or more multipoles contribute to the transition.

Apart from the photodisintegration of the deuteron, the most promising of all the known photonuclear reactions is the reaction $Be^9(\gamma,n)Be^8$. Since the threshold energy is about 1.66 Mev and the cross section at 9.5 Mev is about 1.6 mb, this reaction could compete with the photodisintegration of the deuteron as an effective analyzer of polarization for gamma rays if the value of α were large enough.

(*c*) *Photofission*

Winhold and Halpern (Wi 56), Katz *et al.* (Ka 58), and others have shown experimentally that certain nuclei subject to photofission exhibit anisotropy in the angular distribution of the fission fragments. Two such nuclei, the even-even nuclei, Th^{232} and U^{238}, for example, show angular distributions of the form $(a+b\sin^2\theta)$. Although the form of the distribution remains the same over the energy range studied (5 to 20 Mev), the ratio b/a varies from values of the order of 10 near threshold (just above 5 Mev) to values of the order of 0.1 in the vicinity of 20 Mev. The anisotropic distribution is assumed to result from an electric dipole interaction. Consequently, although the azimuthal distribution produced by polarized photons has not yet been studied, the foregoing angular distribution implies a polarization sensitivity proportional to b/a. Other nuclei, such as the odd-A nuclei, U^{235} and Pu^{239}, have essentially isotropic angular distributions and thus are not expected to manifest any sensitivity to polarization.

Although the photofission cross sections in the energy region of interest here are of the order of 10 mb, the fact that appreciable anisotropy in the angular distribution is confined to the region roughly 2 to 3 Mev above threshold places some limitation on the potential use of photofission as a polarization analyzer. In addition, if neutron-induced fission or spontaneous fission is present, the detection of the photofission might become very difficult.

III. MEASUREMENTS OF POLARIZATION OF NUCLEAR GAMMA RAYS

The following discussion is organized primarily on the basis of the type of polarization-sensitive mechanism involved. Attention has been given to chronological order only insofar as it lends itself to a logical development. The development of technique has also been a consideration in selecting experiments for discussion. A complete listing and outline of the experiments is attempted in Table XIII at the end of Part III. Since this discussion is primarily devoted to experiments on *nuclear* gamma rays, the polarization experiments on annihilation radiation and bremsstrahlung are not discussed except for the earliest annihilation experiments.

A. Measurement of Plane Polarization by Means of the Compton Effect

Not only the earliest, but also the most utilized, of the polarization sensitive reactions has been the Compton effect. This is due to the fact (Part II) that the reaction has a relatively high cross section along with a reasonable sensitivity to polarization over an energy range which includes many nuclear gamma rays of interest.

In general, a polarimeter that uses the Compton effect consists of a scatterer and an analyzer which detects the scattered radiation. The elements of such a polarimeter are shown in Fig. 14. Here the polarimeter consists of the elements B and C. Gamma rays from the source impinge on the Compton scatterer B and the scattered radiation is detected by the analyzing element

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

C. As a general rule B and C are crystals of scintillation counters and coincidence between B and C are measured. This makes it considerably easier to isolate and detect the desired Compton events. A polarization measurement is made by determining N_0/N_{90} (often called $N_{\parallel}/N_{\perp}$), the ratio of the coincidence counting rate when C is placed at $\phi=0°$ to that when it is placed at 90°. In principle, B could be merely a piece of scattering material, and the polarization measurement would depend only on the single counting rate in C at the 0° and 90° positions. However, in this case the shielding (and electronic discrimination) would have to be exceedingly good to prevent direct radiation from the source being recorded in C. See Sec. (a).

Except in the earliest experiments, a polarimeter has generally consisted of an organic scintillator for the scatterer and a sodium-iodide crystal for the detector of the scattered gamma ray. An organic scintillator, a material of low Z, makes a good scatterer because it permits fairly free escape of the scattered photons. Naturally, sodium iodide is the best analyzing crystal because of its high efficiency and good resolution for the detection of gamma rays. The objection to the use of sodium iodide for this purpose in the earliest experiments was the long decay time of its pulses. Development of suitable electronic techniques has greatly diminished this difficulty.

Circular polarization experiments utilizing the Compton effect are discussed in the last section of this part.

(*a*) *Early Experiments on Annihilation Radiation*

Although this paper deals mainly with the polarization of gamma rays of nuclear origin, for convenience and historical interest we begin with a brief discussion of the first measurements of polarization, those performed in order to detect the mutually perpendicular polarizations of the two quanta resulting from the annihilation of a positron and an electron. This prediction of the pair theory was pointed out by Wheeler (Wh 46). Assuming that two Compton polarimeters would be used to detect the polarizations, Pryce and Ward (Pr 47) and Snyder *et al.* (Sn 48) calculated the expected azimuthal distribution of one scattered quantum relative to the other. This is an example of a polarization-polarization correlation with the angle between the two polarimeters fixed at 180°.

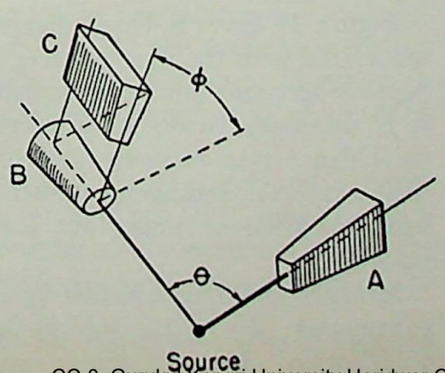


FIG. 14. Schematic diagram of experimental arrangement of scintillating crystals in an apparatus designed to measure a direction-polarization correlation. Crystals B and C constitute the polarimeter; crystal A is the directional counter (figure from Me 50).

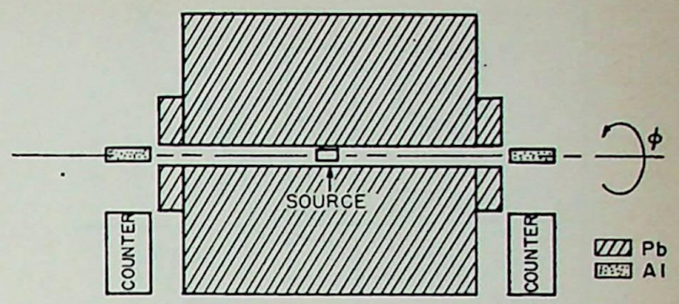


FIG. 15. Schematic diagram of typical experimental arrangement for study of linear polarization of annihilation radiation. In this illustration the scatterers are made of aluminum. The counters are in the parallel (‖) position.

The essential features of the experimental arrangement are shown in Fig. 15. The positron source (Na^{22} or Cu^{64}) is placed at the center of a hole about $\frac{3}{8}$ in. in diameter bored through the center of a lead block with dimensions approximately 6×6×6 in. The oppositely directed annihilation radiation beams emerging from the openings then impinge upon cylindrical scatterers (not detectors in this case) about 1 in. long. The detectors of the scattered radiation are placed so that the mean scattering angle is about 82°, the angle at which the Compton effect exhibits optimum sensitivity to polarization for this gamma-ray energy. (Fig. 2.) The experiment then consists in determining $N_{\perp}/N_{\parallel}$, the ratio of the coincidence counting rate when the axes of the two detectors are at right angles to each other to that when they are parallel.

Bleuler and Bradt (Bl 48) using G-M counters with end-windows as detectors confirmed the predicted correlation, although their results have a relatively large margin of error. They obtained $N_{\perp}/N_{\parallel}=1.9\pm0.3$ as compared to the theoretically predicted ratio of ~1.7 for their geometry. Shortly thereafter, Hanna (Ha 48a) performed the experiment using G-M tubes and obtained values of $N_{\perp}/N_{\parallel}$ which were consistently lower than the theoretical values. Vlasov and Dzhelepov (Vl 49) obtained 1.7 ± 0.2. However, the approximations used in obtaining the theoretically expected ratio for their geometry were sufficiently rough that essentially there was only qualitative agreement with the theory. Wu and Shaknov (Wu 50) repeated the experiment using anthracene scintillation counters and obtained a value of $N_{\perp}/N_{\parallel}=2.04\pm0.08$, which compares well with the theoretical value of 2.00 for their geometry.

(*b*) *Direction-Polarization Correlation*

The arrangement of the scintillation crystals for a direction-polarization correlation experiment is shown in Fig. 14, in which A is the crystal of the counter which is insensitive to polarization and records only the direction of one gamma ray and B and C are the crystals of the Compton polarimeter which measures the linear polarization of the other gamma ray. The phototubes associated with the three crystals are em-

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

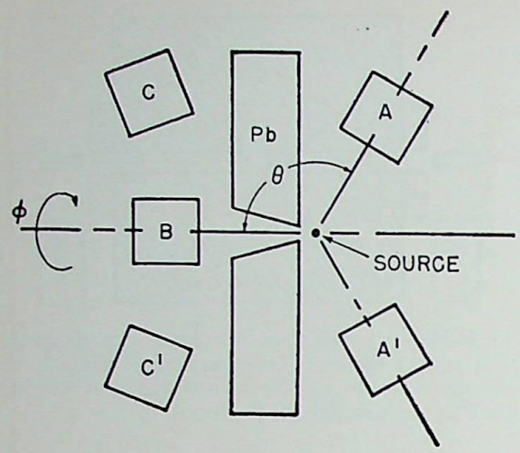


FIG. 16. Schematic diagram of apparatus of Metzger and Deutsch. Two additional counters are used to increase the yield in the measurement of the direction-polarization correlation. In the position shown, $\phi=0$ (after Me 50).

ployed in a triple-coincidence circuit. Thus, a measurement of the polarization is obtained from the ratio of the triple-coincidence yields with crystal C at $\phi=0°$ and at $\phi=90°$. The direction-polarization correlation is then obtained by determining this ratio as a function of the correlation angle θ. The ratio of the counting rates is N_0/N_{90} from which one obtains the true polarization ratio J_0/J_{90} discussed in Part I.

Thorough consideration must be given to the possibility of recording "false" coincidences as a result of radiation being counted and scattered among the three crystals in combinations different from the one desired. Usually a good shield is placed in front of counter C to diminish the amount of direct radiation from the source or scattered radiation from A arriving at C. This shield diminishes the chance that one gamma ray goes to A while the other one goes to C and is then scattered to B, or that one gamma ray goes to B while the other goes to A and is scattered to C.

As mentioned in Part I, there is little to be gained (Ha 48) by putting a polarimeter at A as well as at B, C. Furthermore, if the corresponding direction-direction correlation is known it is only necessary in virtually all cases to measure the direction-polarization correlation at a single angle. This angle is usually selected so as to maximize the effect and it is often 90°.

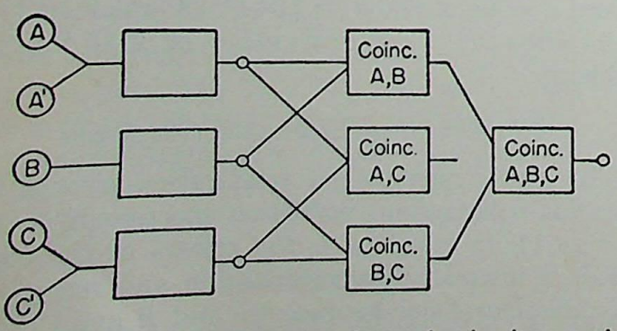


FIG. 17. Block diagram of the electronic circuits associated with the experimental arrangement shown in Fig. 16. Empty blocks represent appropriate stages of amplification, etc. (figure from Me 50).

Metzger and Deutsch (Me 50) first used the technique of the direction-polarization correlation in the study of several gamma-gamma cascades. The method has been extended by others to the study of correlations in which the radiation whose polarization is not analyzed is some other type of particle, either emitted or absorbed.

(c) *Experiments of Metzger and Deutsch*

The classic experiments of Metzger and Deutsch (Me 50) exemplify the techniques involved in the measurement of direction-polarization correlations. Figure 16 is a schematic diagram of the experimental arrangement. Naphthalene was used for all of the crystals in this arrangement, but Metzger and Deutsch suggest the use of anthracene would be an improvement.

Counters A' and C' add to the original coincidence combination ABC three more equivalent (theoretically and experimentally) combinations ABC', $A'BC$, $A'BC'$, so that the real coincidence yield is increased by a factor of four. The addition of C' also, to a first approximation, cancels the effect of any small asymmetry resulting from a possible misalignment of the axis of the polarimeter.

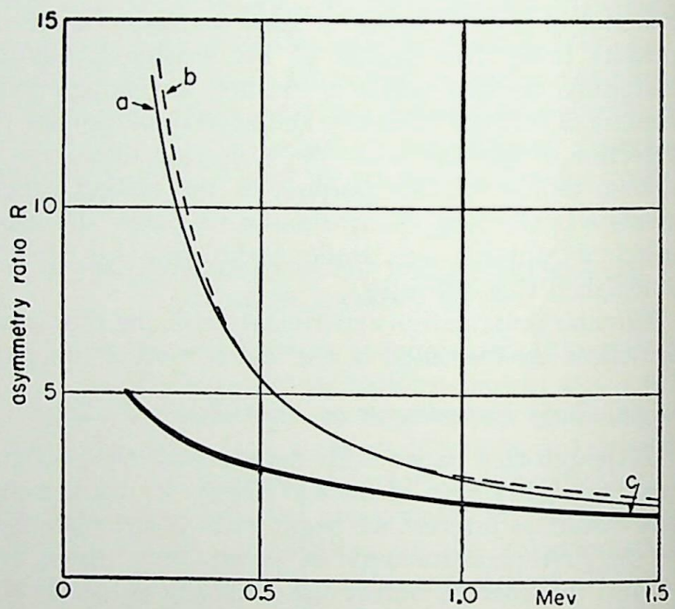


FIG. 18. Asymmetry ratio R as a function of gamma-ray energy in Mev for different geometries: (a) $\theta_c=80°$, ideal geometry; (b) $\theta_c=(\theta_c)$ max, ideal geometry; (c) $\theta_c=80°$, $\Delta\theta_c=55°$, $\Delta\phi=60°$ (figure from Me 50).

Figure 17 gives a block diagram of the electronics associated with the five counters. This system makes it possible to count double coincidences separately, which it is necessary to do in order to monitor the triple-coincidence rate properly and to determine the accidental rate. An arrangement such as this, or its equivalent, has been an important feature of all Compton polarimeters.

Metzger and Deutsch distinguish three polarization parameters: $p=J_0/J_{90}$, which describes the state of polarization of the gamma ray incident upon the

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

polarimeter; R, the asymmetry ratio, which is a measure of the polarization sensitivity of the polarimeter taking into consideration nonideal geometry; and N_0/N_{90}, the ratio of the triple-coincidence counting rates with the polarimeter set at $\phi=0$ and $\phi=90°$. These three parameters are related through the expression

$$N_0/N_{90}=(p+R)/(pR+1). \qquad \text{(III-1)}$$

For the case of ideal geometry (Part II) R is simply the ratio of two differential cross sections for Compton scattering, i.e., $R=d\sigma_{90}/d\sigma_0$. There is an angle of scattering*** $\theta_c=(\theta_c)_{\max}$ at which R is a maximum for ideal geometry. The dependence of R on the energy of the initial photon, as calculated by Metzger and Deutsch, is given as curve b in Fig. 18 for $\theta_c=(\theta_c)_{\max}$. Curve a shows the dependence of R for $\theta_c=80°$, which Metzger and Deutsch chose as a mean value. In order

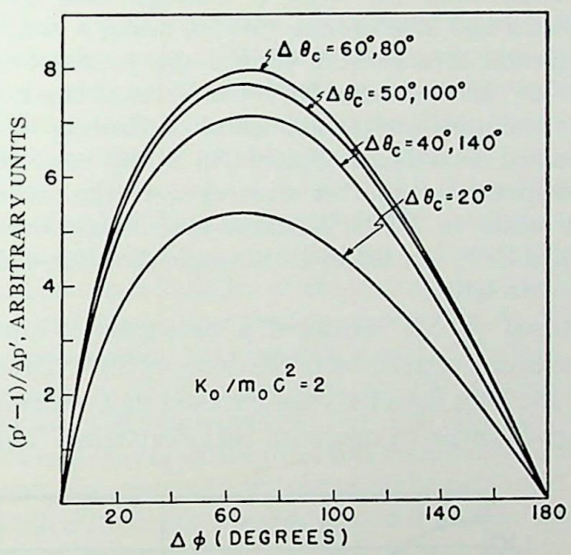


FIG. 19. Figure of merit of a Compton polarimeter as a function of the spread in the angle ϕ, for various spreads in θ_c. Incident photon energy equal to 1 Mev, and $p=1.2$. The curves do not differ significantly from those in Me 50.

to obtain statistically significant counting rates, however, it is necessary to use angular spreads $\Delta\theta_c$ and $\Delta\phi$, which are considerably greater than zero, so that R is appreciably reduced. Metzger and Deutsch use $(p'-1)/\Delta p'$ as a figure of merit for a polarimeter, where p' is the observed polarization ratio, and $\Delta p'$ its error. This quantity is given as a function of the angular spreads in Figs. 19 to 22 for two gamma-ray energies. For angular spreads of 60°–70°, the figure of merit is a maximum over a wide range of energy. Hence, it is possible to design an efficient polarimeter of considerable versatility.

The actual spreads used by Metzger and Deutsch were $\Delta\theta_c=55°$ and $\Delta\phi=60°$. That these were the effective values was determined experimentally in the following way. The partially polarized radiation result-

*** Here, θ_c is used for the scattering angle to distinguish it from the correlation angle θ.

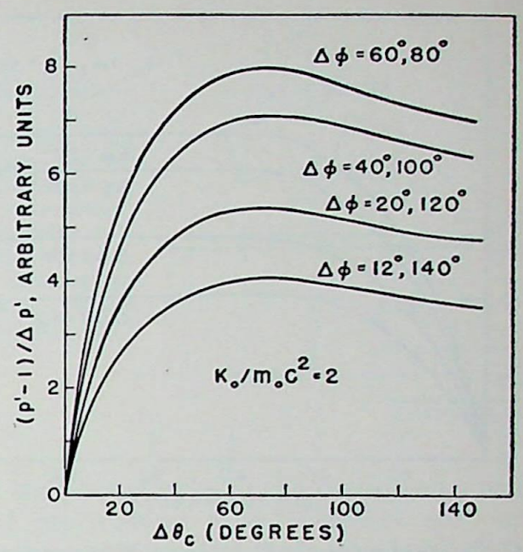


FIG. 20. Figure of merit of a Compton polarimeter as a function of the spread in the angle θ_c, for various spreads in ϕ. Incident photon energy equal to 1 Mev, and $p=1.2$. The curves do not differ significantly from those in Me 50.

ing from the Compton scattering of a collimated beam of gamma rays from a strong Co^{60} source impinged on the scattering crystal of the polarimeter. The degree of polarization p of this radiation is known through the Klein-Nishina formula. A measurement of N_0/N_{90} then determines R through the use of (III-1). With this value of R, the angular spreads $\Delta\theta_c$ and $\Delta\phi$ are obtained for the energy appropriate to the scattered quanta. With these values of $\Delta\theta_c$ and $\Delta\phi$, the curve for R as a function of energy was then constructed and is given as curve c in Fig. 18.

With the equipment and techniques discussed, Metzger and Deutsch measured the direction-polarization correlations in the decays of Sc^{46}, Co^{60}, Rh^{106}, and Cs^{134}. Essentially as a check on the experimentally known direction-direction correlations (Br 48) and to elucidate the method, the polarizations were measured as a function of θ. For each nucleus, possible direction-polarization correlations were calculated from the known coefficients of the direction-direction correlation. Thus, for the cascade $j_1(L_1)j(L_2)j_2\rightarrow 4(2)2(2)0$, the

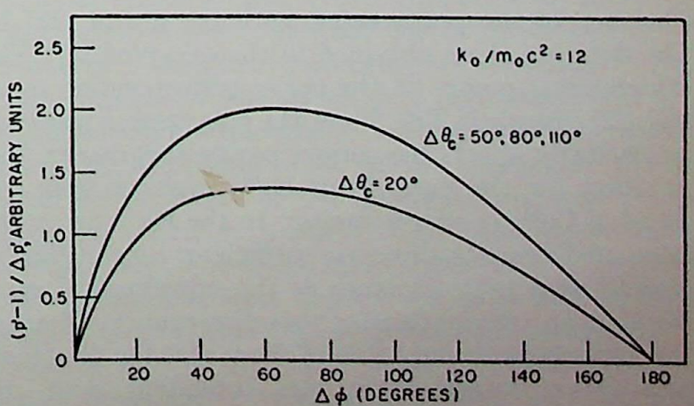


FIG. 21. Figure of merit of a Compton polarimeter as a function of the spread in ϕ, for various spreads in θ_c. Photon energy equal to 6 Mev, and $p=1.2$.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

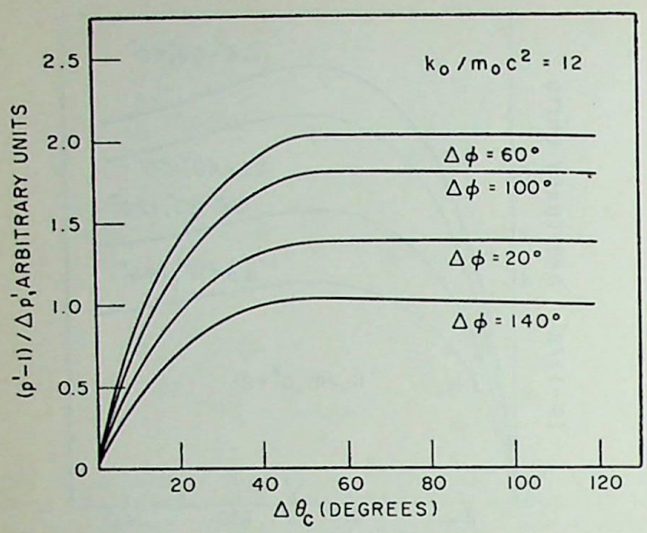


FIG. 22. Figure of merit of a Compton polarimeter as a function of the spread in θ_c, for various spreads in ϕ. Photon energy equal to 6 Mev, and $p=1.2$.

direction-direction correlation is

$$W(\theta)=1+(1/8)\cos^2\theta+(1/24)\cos^4\theta.$$

The direction-polarization correlation can be written down from Table III(b). At $\theta=90°$, for example

$$W(90,\gamma)=1\pm\cos 2\gamma(1/8+1/24).$$

The sign is + or − according as the radiation is electric or magnetic. This expression applies if the first gamma ray goes to the directional counter and the second to the polarimeter, or if the gamma-ray paths are interchanged. In the experiment, both processes were observed indistinguishably with about equal efficiencies. Hence (at $\theta=90°$),

$$p=\frac{J_0}{J_{90}}=\frac{2+(\pm)L_1 0.17+(\pm)L_2 0.17}{2-(\pm)L_1 0.17-(\pm)L_2 0.17}$$

and $p=1.40$, 1, or 0.71 depending on whether the gamma rays in the cascade are $(E2,E2)$, $(E2,M2)$ or $(M2,E2)$, or $(M2,M2)$. The complete correlation can be worked out in similar fashion.

Figure 23 gives the three curves modified by the efficiency of the polarimeter, with the aid of (III-1). The data are those obtained in the decay of Sc^{46}. The experimental points fit the curve corresponding to an $(E2,E2)$ cascade in Ti^{46}. Thus, the polarization measurement clearly selects the correct parity assignment. The situation for the cascade in Ni^{60} resulting from the decay of Co^{60} is entirely similar. In the Rh^{106} decay the calculated direction-direction coefficients for a 0(2)2(2)0 scheme were twice as large as the experimental ones. Without an explanation for this discrepancy, Metzger and Deutsch were only able to show that the direction-polarization measurement was consistent with the direction-direction measurement. The experimental result was interesting in that the polarization changes direction at $\theta\simeq 120°$, attaining large values on either side of this angle. In the case of Cs^{134}, the possible participation of other coincident gamma rays in the correlation made a unique assignment impossible. Again the direction-polarization correlation was consistent with the direction-direction correlation, and the result was not unlike that for Sc^{46} and Co^{60}.

(d) *Other Measurements with Radioactive Sources*

Since the Metzger-Deutsch experiments, there have been many other measurements of direction-polarization correlations in a variety of nuclei. The same basic techniques are common to all. Naturally, there have been modifications and improvements to fill the needs of each new experiment. Accordingly, in the review of the remaining experiments using the Compton polarimeter only those techniques which represent relatively distinctive additions to the art are discussed.

Shortly after the work of Metzger and Deutsch, Williams and Wiedenbeck (Wi 50) using a similar experimental arrangement verified the results on Rh^{106} and Co^{60}. In the case of Rh^{106} the factor of two between the theoretical and experimental coefficients could be explained according to Spiers (Sp 50) by assuming the participation of another state close to the second excited state in Pd^{106}. Williams and Wiedenbeck also studied Cs^{134}, but found an isotropic direction-polarization correlation.

Stump (St 52) measured a beta-gamma direction-polarization correlation in the decay of Sb^{124}. The source and the beta detector were enclosed in a vacuum container in order to minimize beta scattering. The beta

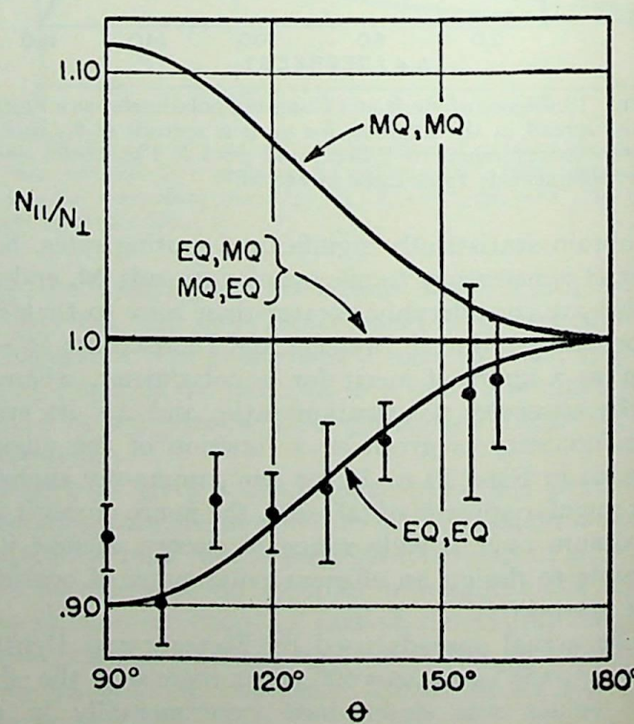


FIG. 23. Gamma-gamma direction-polarization correlation in the decay of Sc^{46}. The ratio $N_{\parallel}/N_{\perp}$ is plotted against θ. The three curves correspond to the different parity assignments in the spin sequence 4-2-0 (figure from Me 50).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

detector consisted of a thin crystal of anthracene mounted on the end of a light pipe leading to a photomultiplier tube outside the vacuum chamber. The polarimeter was considerably improved in efficiency by using an anthracene crystal as the scattering crystal and two oppositely placed sodium-iodide crystals as detectors of the scattered gamma rays. Since the angular correlation coefficients were already known (St 51) and quadrupole radiation is involved, it was necessary to take measurements only with the polarimeter at 90° to the beta counter. Stump found a value of about 1.1 for the ratio N_0/N_{90}. This datum, used in conjunction with the correlation coefficient, led to the conclusion that no parity change occurred in the gamma transition.

Further work on the radioactive nuclei Co^{60}, Cs^{134}, and Sb^{124} was done by Kloepper *et al.* (Kl 52). Plastic scintillators and type 5819 phototubes were used. The phototubes were shielded from magnetic fields in order to reduce the spurious dependence of the coincidence rate upon counter position. Residual effects were investigated by comparing the triple-coincidence rate to the single-channel counting rates at the various positions. In addition to monitoring the single rates, the various double-coincidence combinations, and the resolving time, the instrument was checked with observations at $\theta=\pi$, where N_0/N_{90} must equal 1.0. The gamma-gamma direction-polarization correlation for Co^{60} was measured to establish the reliability of the instrument. The result agreed with the earlier measurements (Me 50, Wi 50).

The polarization measurements on Cs^{134} agreed with the results of Metzger and Deutsch (Me 50), but not with those of Williams and Wiedenbeck (Wi 50). It was pointed out that both the direction-polarization and the direction-direction correlations could be explained with a $4(E2)2(E2)0$ decay scheme, provided a 35% isotropic component is assumed to be present in both correlations.

In the case of Sb^{124} the gamma-gamma directional and polarization correlations were essentially isotropic. The authors suggested a $3(10\%\ E2;\ 90\%\ M1)2(E2)0$ decay scheme. The beta-gamma direction-polarization correlation was also studied with this source. The measurement agreed with the result of Stump (St 52) confirming the $E2$ nature of the radiation.

Additional work on the decay of Cs^{134} was done by Robinson and Madansky (Ro 52) who measured a polarization-polarization correlation for the gamma rays. The experiment was quite similar to the investigations on annihilation radiation (see Fig. 15). In fact, Robinson and Madansky first performed the experiment with annihilation radiation in order to determine experimentally the effective solid angle corrections. The signals from the two scattering crystals (stilbene) went to a coincidence circuit (7×10^{-8} sec), the output of which triggered an oscilloscope. The pulses from the two analyzing crystals (NaI) were delayed, one more than the other, and displayed on the (triggered) oscilloscope. A coincidence analysis could be made by visual inspection of the photographs of the oscilloscope traces. The ratio of the number of coincidences with analyzing counters perpendicular to the number with the counters parallel was 0.91 ± 0.08. This result is not in disagreement with assignments of spins and parities made from the other correlation experiments on this nucleus (Wa 50, Pe 51, Ro 51).

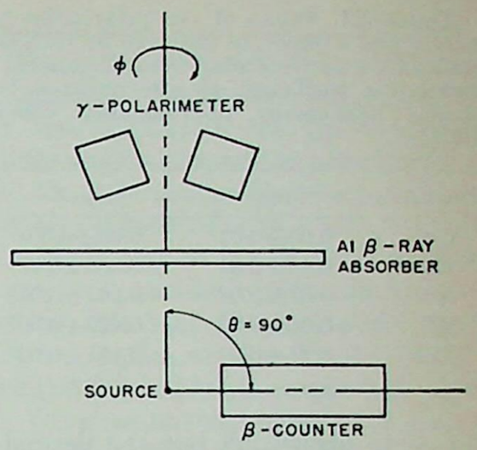


FIG. 24. Schematic diagram of the apparatus of Hamilton *et al.* which uses two equivalent counters in the polarimeter. This apparatus was used to measure beta-gamma polarization correlations (figure from Ha 53a).

A study of the direction-polarization correlation for the cascade involving the first two excited states of Pb^{208} in the ThC″ decay was made by Kraushaar and Goldhaber (Kr 53). Again, a polarimeter of the Metzger-Deutsch type was used and a value for the efficiency was obtained by observing the polarization of scattered Co^{60} radiation. However, a more realistic value for the efficiency of the polarimeter was obtained by measuring the already known direction-polarization correlations for Co^{60} and Rh^{106} sources. For the measurement in the Pb^{208} cascade, a value of $N_0/N_{90}=0.958\pm0.3$ was obtained. This result, in conjunction with the result of Petch and Johns (Pe 50) on the direction-direction correlation, indicates an assignment of $4(E2)2(E2)0$ to the cascade in Pb^{208}.

Direction-polarization correlations in the beta-gamma decay from K^{42}, As^{76}, Rb^{86}, Sb^{124}, and Cs^{134} were studied by Hamilton *et al.* (Ha 53a). They used a two-counter polarimeter having stilbene crystals placed symmetrically with respect to the polarimeter axis, as shown in Fig. 24, so that each counter played a dual role as scatterer and analyzer. The aluminum plate prevented beta particles from reaching the counters of the polarimeter. The entire apparatus was surrounded by an aluminum can which served as a light shield and contained a helium atmosphere in order to reduce electron scattering. Hamilton *et al.* discuss thoroughly the dependence on angle of the spurious triple-coincidence rates associated with the various possible scattering combinations. The single and double rates, as in previous experiments, were determined quite accurately as a check on the proper operation of the instrument.

In the case of Sb^{124}, Cs^{134}, and As^{76} there are additional beta-gamma cascades involving beta particles whose energies are lower than in the cascades under study. In these cases an aluminum absorber was intro-

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

TABLE XI. Values of the polarization correlation coefficient C(obs) and C(pred), as observed by Ha 53a and as predicted (on basis of no parity change for the gamma) from the directional correlation coefficient Q_2+Q_4 reported by other workers for similar beta-energy discrimination conditions. (Table from Ha 53a).

Nucleus	Q_2+Q_4	C(pred)	C(obs)*
Cs^{134}	0.00±0.01	0.00±0.003	0.009±0.013
Rb^{86}	0.13±0.01	−0.068±0.005	−0.060±0.025
As^{76}	0.07±0.02	−0.067±0.018	−0.114±0.035
K^{42}	−0.06±0.02	0.022±0.007	−0.007±0.016
Sb^{124}	−0.27±0.04	0.151±0.022	0.142±0.010

duced to absorb all but the desired high-energy beta particles. The effect of spurious triple coincidences caused by gamma rays reaching the beta crystal was determined by making runs with absorbers of various thickness, but all thick enough to absorb all beta particles.

In these experiments a quantity C is defined which is the coefficient characterizing the angular variation of the intensity of triple coincidences as a function of azimuthal angle of the polarimeter. Thus

$$N_\phi = 1 + C\cos^2\phi, \qquad \text{(III-2)}$$

so that $N_0/N_{90}=1+C$. Table XI compares the values of C obtained experimentally with values derived from the direction-direction coefficients determined by other workers.††† At $\theta=90°$, the polarization ratio is [see Sec. (c)]

$$p=\frac{J_0}{J_{90}}=\frac{1\pm(Q_2+Q_4)}{1\mp(Q_2+Q_4)}$$

for pure dipole ($Q_4=0$) or pure quadrupole radiation, the upper signs holding with no parity change, the lower ones with parity change. From (III-1), one obtains N_0/N_{90} and hence C. In the case of Cs^{134}, an isotropic direction-polarization correlation is observed as predicted. For Rb^{86}, As^{76} and Sb^{124}, the observed polarization agrees with that predicted for no parity change in the gamma transition from the first excited state. An essentially isotropic direction-polarization correlation is observed for K^{42}.

Brazos and Steffen (Br 56) have studied the direction-polarization correlation of the 0.722–0.566-Mev gamma-ray cascade resulting from the decay of the 50-day In^{114} isomer to Cd^{114}. The oppositely placed analyzing crystals were rotated automatically about an axis through the scatterer and source once every hour to minimize the effect of possible electronic fluctuations. The measurement indicated a $4(E2)2(E2)0$ cascade, which is also favored by the directional correlation measurements on this cascade (Br 56).

The direction-polarization correlation in the 1.24–0.845-Mev gamma-ray cascade in the decay of Co^{56} has been measured by Wood and Jastram (Wo 55). A liquid scintillator was used as the scattering counter. Using the direction-direction coefficients (Hu 55) and the polarization measurement on the 0.845-Mev gamma ray, it was established that this gamma ray was electric quadrupole (no parity change).

An investigation of the spins and parities of the first three excited levels in Ce^{140} and the first two in Sr^{88} was conducted by Bishop and Perez y Jorba (Bi 55) by means of direction-direction and direction-polarization correlations. They used scattering crystals (made of polystyrene activated with diphenyl tetrabutadiene) of two sizes, one for low the other for high-energy gamma rays, since too large a crystal at a given energy makes double scattering too probable. They limited the double scattering to 15% in their experiments. Single-channel pulse-height analyzers were used in the channels of both the directional counter and the scattering counter. In the case of the scattering counter, energy discrimination is desirable since the polarimeter is in effect a crude Compton spectrometer. Thus, the window of the pulse-height analyzer selects the pulses from Compton recoil electrons in the scattering crystal corresponding to a gamma ray of a given energy scattered at a given (approximately) angle. Figure 25 shows a pulse-height spectrum obtained this way for the gamma rays from Sr^{88} following the beta decay of Y^{88}.

The decay scheme in Ce^{140}, following beta decay of La^{140}, is fairly complex as shown by the level diagram in Table XIII. Bishop and Perez y Jorba measured the direction-polarization correlation with the 1.60-Mev gamma ray going to the directional counter and either the 0.329-, 0.487-, or 0.815-Mev gamma ray going to the polarimeter. The experimental value of N_0/N_{90} (1.077±0.022) was consistent with assignments of $3^--4^+-2^+$ or $3^+-4^--2^-$ for the first three excited states taken in descending order. However, on the basis of the Goldhaber-Sunyar rule, giving 2^+ for the first excited state, and on considering transition probabilities in the beta decay to these states, the first possibility was naturally chosen. Again, in the case of the simple cascade in Sr^{88} the measurements were most consistent with the possibilities $3^--2^+-0^+$ or $3^--1^+-0^+$ for the first two excited states and the ground state. The first possibility was chosen since it agrees best with internal-conversion data (Pe 48, Me 52a).

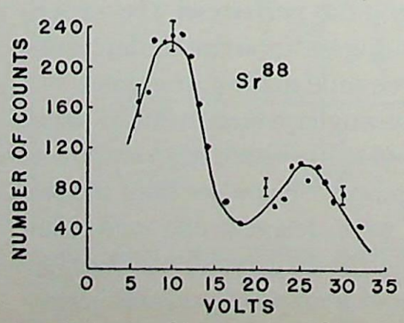


FIG. 25. Pulse-height spectrum of recoil electrons (in the scatterer) which are in coincidence with scattered gamma rays detected by the analyzing counters of the polarimeter. The gamma rays are from Sr^{88} produced in the decay of Y^{88} (figure from Bi 55).

††† The directional correlation results used were the following: K^{42} (Be 50b, St 51), As^{76} (Wa 50a, Ri 52), Rb^{86} (St 51, Ri 52), Sb^{124} (Be50b), Cs^{134} (Be 50b, St 50).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

The Sr^{88} gamma rays were also studied by Coleman (Co 56). His Compton polarimeter was used as a pulse-height analyzer by summing the pulse from the scattered electron in the scattering scintillator with that from the scattered gamma in the analyzing counter. With this technique the analyzing counter can subtend a large solid angle at the scattering scintillator without loss of resolution. The pulse-height spectrum obtained with this "total absorption" polarimeter is presented in Fig. 26. Separate polarization measurements were made on each member of the 0.91–1.85-Mev cascade in Sr^{88}. The results show that the parities of the 1.85- and 2.75-Mev levels are even and odd, respectively, in agreement with the results of Bishop and Perez y Jorba (Bi 55).

Later, using the same equipment, Coleman (Co 56a) studied the polarization correlation of the 0.99–1.33-Mev gamma-ray cascade in Ti^{48} produced in the decay of V^{48}. He discusses the problem of ambiguity arising from a multipole mixture when the result of a direction-direction correlation is used to determine level assignments. Brazos and Steffen (Br 56) showed in their work on the Cd^{114} levels that the ambiguity could be reduced by measuring the direction-polarization correlation. In addition, the mixing ratio may be determined when both correlations are known. This problem is illustrated in Part I. For the cascade in Ti^{48}, pure quadrupole radiation was established giving the assignment $4(E2)2(E2)0$.

The gamma-gamma direction-polarization correlations in the Co^{60} and Na^{24} decays were studied by Estulin *et al.* (Es 56). The experimental arrangement used was similar to that of Metzger and Deutsch. The measurements on Co^{60} were made to determine the reliability of those on Na^{24}. In each case the value of N_0/N_{90} indicated an $(E2,E2)$ transition in the cascade, leading to assignments of positive parity for the first two excited states in Ni^{60} and Mg^{24}.

Stelson and McGowan (St 57) have measured the direction-polarization correlation in the 133–482-kev cascade in Ta^{181}. They used NaI(Tl) crystals, 3 in. thick and 3 in. in diameter, in the counters of the correlation arrangement except in the Compton scatterer which was an anthracene scintillator, $1\frac{1}{2}$ in. thick by $\frac{3}{4}$ in. in diameter. The output of a fast-slow triple-coincidence circuit gated a 20-channel pulse-height analyzer on which was displayed the pulse-height spectrum from the analyzing counter of the polarimeter. The effectiveness of the polarimeter had been determined by measuring the polarization of gamma rays resulting from the Coulomb excitation of 2^+ levels in even-even nuclei (see next section). The direction-direction measurements (Mc 54a, Pa 55, He 55) had indicated two possible assignments for the 133–482-kev decay: $5/2(E2)9/2(E2+M1)7/2$ and $1/2(E2) \times 5/2(E2+M1)7/2$. The experimental value of N_0/N_{90} obtained in the polarization measurement clearly agreed with the latter assignment with even parities for the levels. The direction-polarization correlation was measured with both liquid and solid Hf^{181} sources as a check on possible systematic errors. It was estimated that the 4% contribution of 136.82-kev radiation from the decay of the 618.9-kev state just above the 615-kev state, did not significantly affect the results.

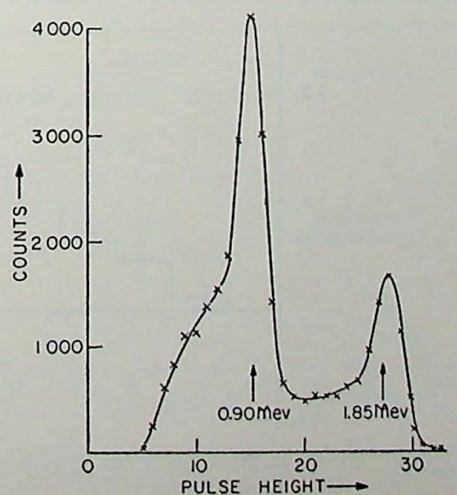


FIG. 26. Pulse-height spectrum of gamma rays from the same decay as in Fig. 25, observed using a "total absorption" polarimeter (figure from Co 56).

(e) *Measurements on Gamma Rays from Nuclear Reactions and Coulomb Excitation*

French and Newton (Fr 52) applied these methods to an alpha-gamma direction-polarization correlation in the $F^{19}(p,\alpha)O^{16*}(\gamma)O^{16}$ reaction. This experiment was exceptional in that an octupole transition (Ar 50, Ba 50) is involved and complete polarization is expected at an angle of 118° with respect to the coincident α particle as seen in Part I. Furthermore the gamma-ray energy was 6.13 Mev. Even though both the crystal absorption and the polarization sensitivity of the Compton process are low at this gamma-ray energy, a Compton polarimeter was still used successfully. Figure 2 shows that $R=1.27$ at this energy. After correcting for the angular spreads in the apparatus, the authors obtain values of 0.84 and 1.19 for the asymmetry ratio for magnetic and electric octupole radiation, respectively. Since the experimental ratio was 1.14 ± 0.06, it was concluded that the radiation is electric octupole. This gave an assignment of odd parity to the 6.13-Mev state in O^{16}, and even parity to the compound state in Ne^{20} and the ground state of F^{19}, assuming that the ground state of O^{16} is even.

McGowan and Stelson (Mc 58) have measured the polarization of gamma rays resulting from the Coulomb excitation of several odd-A nuclei. In general, these gamma-rays are from mixed $M1+E2$ transitions. Sometimes, a measurement of the direction-direction correlation between the gamma rays and the incident beam will determine the ratio $E2/M1$ in addition to specifying the spin of the excited state which is involved. However, the value of $E2/M1$ or of the spin often remains ambiguous. Such ambiguities can usually be resolved by a direction-polarization measurement.

The equipment and the techniques were essentially

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

the same as those used by these workers in their study of the gamma rays from Ta^{181} discussed in the preceding section. The only basic change is that the bombarding beam and target replace the directional counter and source, since the correlation in this case is between bombarding particle and gamma ray. With this method ambiguities in $E2/M1$ for the gamma-ray transitions or uncertainties in the spin of the first excited states of the following nuclei have been resolved: Mo^{95}, Rh^{103}, Ag^{107}, Ag^{109}, Cd^{111}, Cd^{113}, Au^{197}, Tl^{203}, and Tl^{205} (see Table XIII).

Litherland and Gove (Li 58) have measured the polarizations of the 1.64-, 0.94-, and 1.37-Mev gamma rays from the $C^{12}(He^3,p)N^{14}$, $O^{16}(He^3,p)F^{18}$, and $Mg^{24}(p,p'\gamma)$ reactions, respectively. The measurements on the 1.64-Mev gamma ray showed that the 3.95-Mev level in N^{14} had the same parity (positive) as the first excited state. Similarly the 0.94-Mev level in F^{18} had the same parity (positive) as the ground state. The measurements on Mg^{24} were performed at the 2.01- and 2.40-Mev resonances and in both cases very large polarizations were found for the gamma rays emitted at 90°. The results also indicate that the first excited state of Mg^{24} has the same parity (positive) as the ground state.

(*f*) *Measurements on Gamma Rays from Aligned Nuclei*

Of the several possible methods used for orienting nuclei,‡‡‡ the two that have been used the most in the experiments with gamma-ray polarization are the Bleaney method (nuclear alignment) and the Rose-Gorter method (nuclear polarization). Either one or both of these methods was used in the gamma-ray polarization experiments now to be described. The method of Rose and Gorter (Go 48, Ro 49) takes advantage of the nuclear polarization produced by magnetic hyperfine coupling in paramagnetic ions which have been polarized by an external magnetic field. In Bleaney's method (Bl 51a) no external field is used. Instead, the internal crystalline field is used to produce the nuclear alignment, through the magnetic hyperfine coupling in paramagnetic ions. In any nuclear orientation method the very low temperatures which are required are produced by adiabatic demagnetization.

The gamma rays emitted by aligned nuclei are linearly polarized. In fact, all the phenomena and the nomenclature which are associated with the direction-polarization correlation have their counterpart in the alignment-polarization correlation.

Bishop *et al.* (Bi 52), using Bleaney's method, made polarization measurements on the gamma rays emitted from aligned sources of Co^{58} and Co^{60}. After adiabatic demagnetization, values of $N_{\parallel}/N_{\perp}$ were measured with a Compton polarimeter as a function of $1/T^*$ by letting the source warm up (Fig. 27). The magnetic temperature T^* is obtained by measuring the susceptibility and applying Curie's law. The polarization of both gamma rays together in the 4–2–0 cascade in Ni^{60} and of the 805-kev gamma ray from Fe^{58} showed that all these quadrupole transitions (established in the case of Fe^{58} by Da 52) were electric.

Bishop *et al.* (Bi 54) in a later experiment measured the linear polarization of the gamma rays from aligned Mn^{54} nuclei. This was done both by Bleaney's method and by that of Rose and Gorter, a greater gamma-ray polarization being obtained with the latter technique. Since the Rose-Gorter method uses an external magnetic field, the scintillation crystals of the polarimeter were mounted on the ends of light pipes, and the photomultipliers were magnetically shielded. With each method the values of $N_{\parallel}/N_{\perp}$ (obtained as a function of $1/T^*$) showed that the radiation from the first excited state of Cr^{54} at 835 kev had positive parity and was therefore electric quadrupole (Gr 54).

Again using Bleaney's method, Bishop *et al.* (Bi 55a) measured the polarization of the 123-kev radiation from the 137-kev level of Fe^{57} in the decay of Co^{57}. This work was accompanied by a measurement of the alignment-direction correlation. It was known (Al 54) that the 123-kev transition is predominantly dipole. The measurement of $N_{\parallel}/N_{\perp}$ as a function of $1/T^*$ indicated that the transition is predominantly magnetic, thus confirming the suggested parity assignments to the 14- and 137-kev levels (Le 55). From the polarization measurements and the directional correlation a value of $(E2/M1)^{\frac{1}{2}}=+0.19\pm0.02$ was obtained.

Cacho *et al.* (Ca 55), using Bleaney's method, found a value for $N_{\parallel}/N_{\perp}$ of 1.58 at $1/T^*=40$ for radiation from the 145-kev first excited state of Pr^{141} in the Ce^{141} decay. This led to the conclusion that the transition is predominantly magnetic dipole. The polarization study in conjunction with the directional correlation (Ca 55, Am 56) gave $(E2/M1)^{\frac{1}{2}}=0.08\pm0.02$.

The linear polarizations of the quadrupole (Hu 56a) gamma rays, coming from the 6–4–2–0 cascade in Cr^{52}, following the decay of Mn^{52} were investigated by Huiskamp *et al.* (Hu 57) using the Rose-Gorter method. The Compton polarimeter which they used also served as a Compton spectrometer [see Sec. (d)] to analyze the

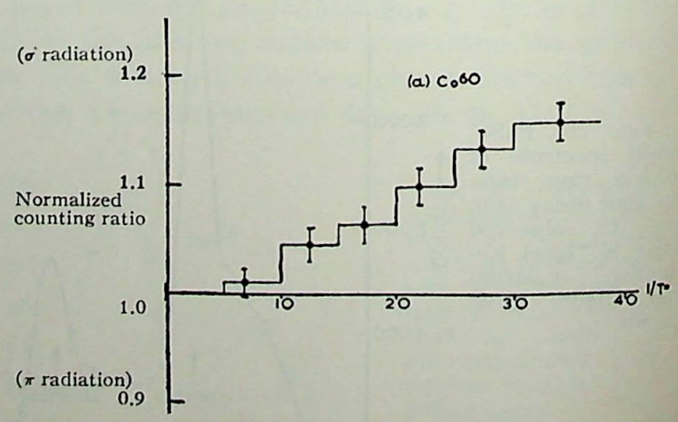


FIG. 27. Linear polarization of the gamma rays from an aligned Co^{60} source as a function of $1/T^*$ (figure from Bi 52).

‡‡‡ Extensive discussions of the methods of orienting nuclei are given in St 57a and Bl 57.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

0.73-, 0.94- and 1.46-Mev gamma rays in the cascade. The energy resolution was sufficient to determine the polarization of each of the three gamma rays. In each case, the radiation was found to be electric.

The same apparatus and method also were used by Diddens *et al.* (Di 58) to polarize Co^{56} nuclei in the study of the levels in Fe^{56}. The Compton polarimeter used as a spectrometer displayed sufficient resolution (14% at 2.76 Mev, 25% at 1.2 Mev) so that the linear polarizations of five gamma rays (0.845, 1.24, 1.75, 2.61, and 3.25 Mev) could be measured satisfactorily. As in the work on Mn^{52} a plastic scintillator was used as the scatterer in the polarimeter instead of NaI(Tl) because it was found that pair production in the NaI with its resulting annihilation radiation produced a serious background for the more energetic gamma rays. The authors also investigated the alignment-direction correlations for the above and other gamma rays produced in the decay of Co^{56}. From all their measurements and with the work of others (Sa 55a) they obtained spins and parities for the 0.845-, 2.08-, 3.45-, 3.84-, and 4.10-Mev levels (see Table XIII). Also the 0.845-, 1.24-, and 3.25-Mev gamma rays were found to be electric quadrupole, whereas the 1.75- and 2.61-Mev gamma rays were magnetic dipole.

B. Measurement of Plane Polarization by Means of the Photodisintegration of the Deuteron

Next to the Compton effect, photodisintegration of the deuteron has proved the most useful polarization-sensitive reaction for gamma rays. In Part II, it was shown that the electric dipole photodisintegration serves as an ideal polarization-sensitive mechanism for gamma rays of energies ranging from about 4 Mev to about 12 Mev. The main disadvantages of the reaction are its low cross section ($\sim 10^{-27}$ cm^2) and the fact that the presence of the isotropic magnetic interaction makes it unsuitable for gamma rays whose energies are much below 4 Mev. It can however be used well above 12 Mev without serious loss of sensitivity.

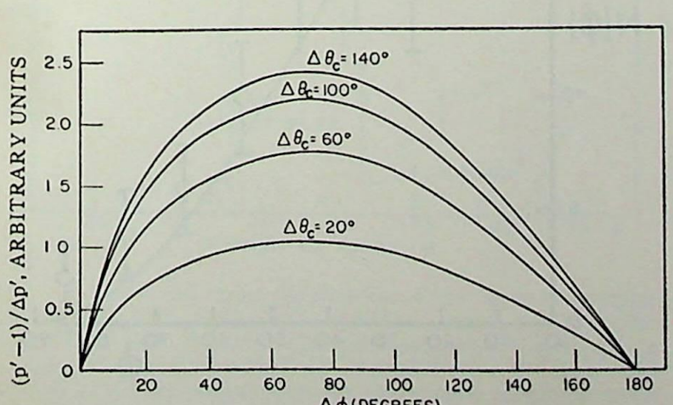


FIG. 28. Figure of merit of a polarimeter using the photodisintegration of the deuteron, plotted as a function of the spread in the angle ϕ, for various spreads in θ_c (computed for $p=1.2$).

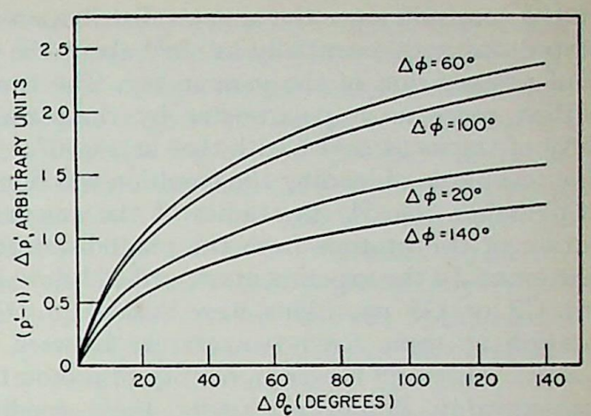


FIG. 29. Figure of merit of a polarimeter using the photodisintegration of the deuteron, plotted as a function of the spread in angle θ_c, for various spreads in ϕ (computed for $p=1.2$).

(a) *Experimental Considerations*

Two methods of utilizing the polarization sensitivity of the deuteron are apparently feasible. One method would be very similar to the Compton polarimeter. The gamma rays would impinge on a "scatterer" containing deuterium, and the photoneutrons would be detected in one (or two) side counters. Almost all the discussion of the Compton polarimeter can be applied to this case. Thus, the advantage of making the "scatterer" itself a detector is obvious. In this case the photoproton would be detected. A very serious drawback to such a polarimeter would be the very high background of events from Compton scattering and pair production. This difficulty might be overcome with modern techniques such as the use of time-of-flight to isolate the neutrons. The advantage of the method over the Compton polarimeter would be its increased sensitivity to polarization at high gamma-ray energy. Its advantage over the emulsion technique, described in the following, would be its possible use in a direction-polarization measurement requiring the detection of coincident radiations. Figures 28 and 29 show the figure of merit for such a polarimeter for various angular spreads. The conditions are the same as in the corresponding calculations for the Compton polarimeter, Figs. 19 to 22, except that here the energy is not an important parameter as long as the electric dipole disintegration is dominant (as assumed in Figs. 28 and 29). That the polarization sensitivity is practically independent of θ_c is reflected in the continued rise of the curves as $\Delta\theta_c$ increases in Fig. 29. The curves in Figs. 28 and 29 can be used in the analysis of data obtained in the emulsion technique.

The simplest polarimeter using the photodisintegration of the deuteron, and the one used in all the experiments performed so far, consists of a nuclear emulsion impregnated with heavy water (D_2O). The emulsions are mounted so that the plane of the emulsion is normal to the direction of propagation of the gamma rays. In this position the maximum number of photoprotons will

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

lie in the emulsion since the angular distribution of the photoprotons goes essentially as $\sin^2\theta$ about the direction of propagation of the gamma ray. The emulsion can then serve as a polarimeter by comparing the density of tracks at $\phi=0°$ with that at $\phi=90°$.

The technique of loading the emulsion was first used by Goldhaber (Go 51) in studies of the angular distribution of the protons from the photodisintegration of deuterons. In the experiments described below, either Ilford C2 or G5 emulsions were soaked in D_2O to saturation at some fixed temperature between 10°C and 20°C. Although the grain density of proton tracks is considerably diminished under these conditions, satisfactory range definition can be obtained in Ilford C2 emulsions with appropriate developing techniques. Better definition can be achieved using Ilford G5 emulsions but in this case precautions must be taken to reduce fogging if low-energy gamma or x-radiation is present.

The amount of water absorbed by a saturated emulsion increases with increasing temperature. However, the emulsion generally becomes too soft to use at temperatures much above 20°C. Saturation of the plates is desirable in order to attain a uniform distribution of water throughout the emulsion. If range determinations are to be made, it is essential for the proton range for a given energy to be constant throughout the emulsion. A uniform distribution of the water is important also in making depth measurements in the emulsion. If the expansion of the emulsion on soaking and the subsequent shrinkage on development do not take place uniformly, it is not possible to obtain accurate measurements on the vertical range of a track. A depth measurement depends on a knowledge of the shrinkage factor (ratio of emulsion thickness before and after development). For most purposes an adequate measurement of the shinkage factor can be obtained with a micrometer, although more elegant techniques are possible. While the emulsions are being exposed they are kept in watertight containers at a controlled temperature. Usually one plate in an exposure is saturated with H_2O to provide a background measurement.

(*b*) *Experiments Using Nuclear Emulsions*

Emulsions soaked in D_2O were first used in a polarization experiment by Wilkinson (Wi 52) in a study of the gamma rays from the $H^2(p,\gamma)He^3$ reaction. In this nonresonant reaction, the energy of the gamma ray is greater than 5.5 Mev by an amount which depends on the energy of the bombarding protons. Since the directional distribution is given by $\sin^2\theta$ (Fo 49), the gamma rays are expected to be essentially completely polarized (Part I).

Ilford C2 emulsions 100 μ thick were soaked in D_2O and placed about 7 cm from a D_3PO target. Although the reaction was weak, 120 usable proton tracks were found in the emulsion after a five-hour exposure with a 50-μa proton beam (1.1 Mev). Wilkinson adopted some simple range and depth criteria to insure the counting of desired events. Since the polarization sensitivity is independent of the polar angle, only the azimuthal angle is recorded. Figure 30 shows the azimuthal distribution of proton tracks projected on the plane of the emulsion. It confirms that the gamma rays are electric dipole and are essentially perfectly polarized. On the other hand, if it is assumed from other evidence (Fo 49) that the gamma rays are completely polarized, then the experiment serves to confirm the ideal polarization sensitivity of the photodisintegration of the deuteron for gamma rays of about 6 Mev.

Fagg and Hanna (Fa 53) used the photodisintegration process to measure the polarization of the 6.13-, 6.9-, and 7.1-Mev gamma rays from the $F^{19}(p,\alpha)O^{16*}(\gamma)O^{16}$ reaction, at $E_p=0.874$ Mev. Ilford C2 emulsions 400 μ thick were soaked to saturation in D_2O and placed 4 cm from a thin target, with the plane of the emulsion normal to the gamma rays emitted at 90°. Several exposures of about 250 μa-hours each were made. Complete resolution of the 6.9-, and 7.1-Mev gamma rays could not be expected but it was essential to isolate the strong 6.13-Mev radiation from the other two. This could be done in a simple manner which permitted (relatively) rapid counting, with the result shown in Fig. 31. In essence, a single depth criterion was adopted: an acceptable proton track must be in sharp focus along its complete range. The utility of the technique is seen by writing the photoproton distribution, (II-14), in terms of projected range $p=R\sin\theta$:

$$d\sigma\sim p^3(R^2-p^2)^{-\frac{1}{2}}dp.$$

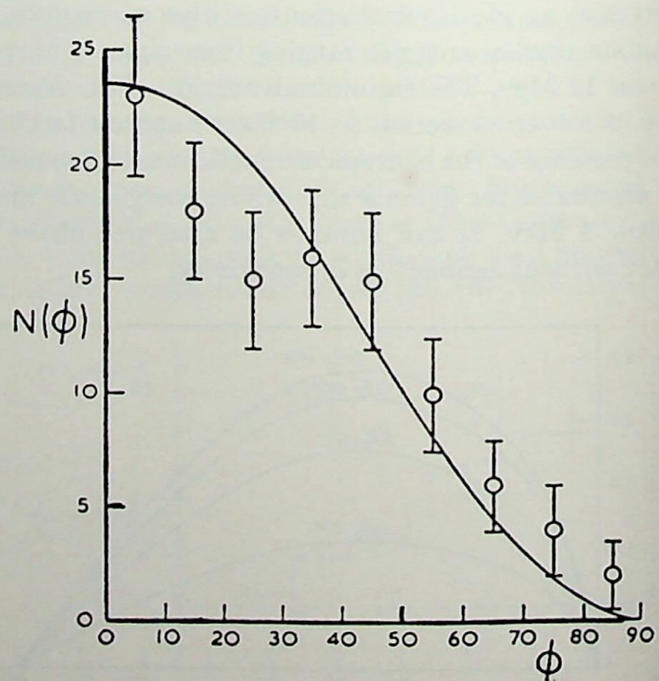


FIG. 30. Linear polarization of gamma rays from $H^2(p,\gamma)He^3$, demonstrated by photodisintegration of deuterons in impregnated emulsion. The azimuthal distribution of photoproton tracks projected onto the plane of the emulsion is shown (figure from Wi 52).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

This distribution is plotted as the dotted line in Fig. 31. Experimentally, it gives a very pronounced peaking at $p=R$. The cutoff at p_c is of course established by the depth criterion. In this case p_c was chosen to resolve the 6.13-Mev gamma ray. The resolution could be improved, by increasing p_c, without greatly decreasing the total number of accepted tracks.

With this technique enough resolution was obtained to give qualitatively the polarizations of the 6.9- and 7.1-Mev gamma rays. This qualitative information was used along with the measured polarization ratio and the directional distribution coefficient (Sa 52) for the unresolved group to deduce the parities of the 6.91- and 7.12-Mev states in O^{16}. These parities (plus and minus, respectively) agreed with those given by Seed and French (Se 52). No assignment could be made for the 6.13-Mev level since the corresponding polarization distribution was essentially isotropic, in agreement with the isotropy observed by Sanders (Sa 52) in his directional distribution measurement.

Nuclear emulsions impregnated with heavy water were also used by Hughes and Sinclair (Hu 56) in their study of the polarization of the gamma radiation from the reactions $Al^{27}(p,\gamma)Si^{28}$, $Mg^{26}(p,\gamma)Al^{27}$, and $Na^{23}(p,\gamma)Mg^{24}$. Ilford G5 emulsions (300 μ and 400 μ thick) were used in order to give well-defined tracks and to avoid fading of the tracks during the long exposures. A copper absorber $\frac{1}{8}$ in. thick was effective in preventing low-energy x-rays from blackening the plates. The targets used were thin enough so that in each bombardment a single resonance was isolated. Both the horizontal and vertical projection of each proton track was measured in the microscope and the range was calculated from these values. The following gamma rays were investigated: the 7.5- and 10.4-Mev gamma rays from Si^{28}, the 7.6- and 8.6-Mev gamma rays from Al^{27}, and the 7.2-, 8.1-, and 10.8-Mev gamma rays from Mg^{24}. A definite polarization was observed in the case of the 10.4-, 7.6-, and 10.8-Mev radiations, all of which were found to be electric in character. The 10.4-, 7.6-, and 10.8-Mev radiations had each been shown to be dipole in character in Ru 54, Ru 56, and Gr 55, respectively. The character of the radiation was used to assign even parity to the ground state and first excited state of Al^{27} (in agreement with Da 53) and to the ground state of Na^{23}. For the remainder of the gamma rays no polarization was observed. Hughes and Sinclair also include in an appendix a range-energy relation for protons up to 5 Mev in wet emulsions.

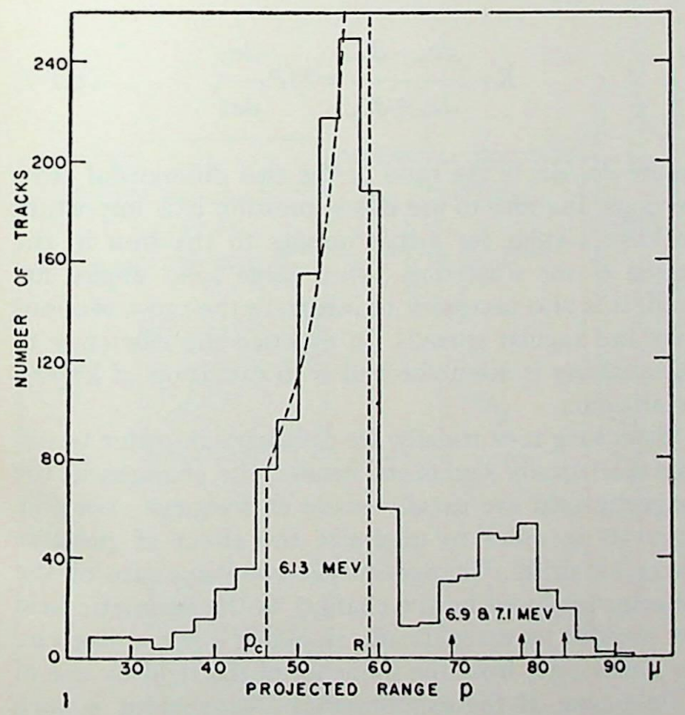


FIG. 31. Projected range distribution of the photoprotons produced by the 6.13-, 6.91-, and 7.12-Mev gamma rays from the $F^{19}(p,\alpha)O^{16*}(\gamma)O^{16}$ reaction. The photoprotons are produced in emulsions impregnated with D_2O. Dashed curve is the theoretical distribution for the projected range, assuming an empirical cutoff at P_c. R is range of photoproton resulting from the 6.13-Mev gamma ray. Arrows in the vicinity of 80 microns mark expected ranges of photoprotons from the 6.91- and 7.12-Mev gamma rays (figure from Fa 53).

C. Measurements of Plane Polarization by Means of the Photoelectric Effect

In Part II it was shown that the usefulness of the photoelectric effect for detection of gamma-ray polarization is rather doubtful. The experiments which have been performed have been primarily motivated by a desire to elucidate the photoelectric effect rather than to measure polarization. With an arrangement very similar to Fig. 15, Hereford (He 51) used two photoelectric polarimeters to observe the cross-polarization of annihilation radiation. In an improved experiment Hereford and Keuper (He 53) substituted a Compton polarimeter for one of the photoelectric detectors. These investigations were followed by the experiments of McMaster and Hereford (Mc 54) and of Brini *et al.* (Br 57) in which the polarization produced in Compton scattering was used to investigate the photoelectric effect, with the discordant results already noted in Part II.

D. Measurement of Circular Polarization by Means of the Compton Effect

Recently, considerable interest has been exhibited in the measurement of circular polarization of gamma rays. This interest has been stimulated primarily by the fundamental reformulation of the law of conservation of parity in weak interactions and the resultant introduction and confirmation of the two-component neutrino theory (Le 56, Le 57). The developments in the measurements of the circular polarization of gamma rays have progressed at a rapid rate. It is the aim here to bring together the methods and techniques of the various investigations and to collect the results which have appeared.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

TABLE XII. Sign of the circular polarization effect R for right circular polarization, for the conditions listed.

Sign of R	Transmission	Scattering
−	$E_\gamma<0.63$ Mev	$\theta_c<90°$
+	$E_\gamma>0.63$ Mev	$\theta_c>90°$

(a) *Experimental Considerations*

A discussion of the Compton scattering of circularly polarized gamma rays by polarized electrons has been given in Part II. The most convenient source of polarized electrons is a ferromagnet. Experimentally the circular polarization of gamma rays is detected by observing the difference in the scattering from electrons in magnetized iron when the magnetization is reversed or when it is removed. This may be done by observing (1) the gamma rays transmitted through the magnetized iron, (2) those scattered from the iron at same appropriate angle, or (3) the electrons ejected forward from the surface of the iron. Although a considerable effect (change in the counting rate with change in magnetic field) could be expected for perfectly polarized gamma rays and electrons at certain angles and energies (Figs. 4 and 5), in practice the effect is greatly diminished by the fact that only about 2 electrons of the 26 in each iron atom can be polarized by magnetization. Therefore, the polarization effect is at most only a few percent; and quite often it is less than one percent, if the gamma-ray polarization is small.

For any one of the above methods, let N_+ denote the experimental yield of gamma rays when the magnetization of the analyzer is directed toward the source, and N_- the yield when it is directed away from the source. The experimental effect is given by

$$R=2\frac{N_+-N_-}{N_++N_-}. \qquad \text{(III-3)}$$

The sign of the effect is of great importance. It is determined by the sign of the polarization-sensitive cross section (Figs. 4 and 5) and the sense of the photon polarization. For ready reference the sign of R is given in Table XII for right circular polarization (see definition in Part I, C) and for four experimental conditions. The sign of R is reversed for left circular polarization. Hence, from the sign of the effect, one may determine the sense of the circular polarization and so also of the polarized emitting nucleus. This much alone has led to very important discoveries.

In accordance with the theoretical definition (Part I, C), the degree of circular polarization is

$$P_3=\frac{J_r-J_l}{J_r+J_l} \qquad \text{(III-4)}$$

where J_r and J_l are the number of right and left circularly polarized photons. Let J_r^0 and J_l^0 be the initial intensities and $J_r^\pm$ and $J_l^\pm$ the transmitted intensities corresponding to (+) or (−) magnetization. Then

$$J_r^\pm=J_r^0\exp\{-nt(\sigma_0 z\pm\sigma_1\nu)\}$$
$$J_l^\pm=J_l^0\exp\{-nt(\sigma_0 z\mp\sigma_1\nu)\}.$$

With $N_\pm=J_r^\pm+J_l^\pm$, we obtain for (III-3)

$$R=2P_3\tanh(-nt\sigma_1\nu), \qquad \text{(III-5)}$$

where n is the number of iron atoms per unit volume, t is the effective thickness of the iron analyzer, and ν is the number of polarized electrons per iron atom, equal to 2.06 at saturation (Ar 53). In using this formula to obtain P_3 from a measurement of R, it is necessary to find appropriate values for t and ν. Alternatively, the analyzer may be calibrated with a source of known polarization.

For the differential cross section we can write

$$d\sigma=d\sigma_0\pm fP_3 d\sigma_1, \qquad \text{(III-6)}$$

where $d\sigma_1$ is the polarization-sensitive cross section shown in Fig. 4 for the case in which the electron polarization is parallel to the photon momentum. The circular polarization P_3 and the electron polarization f are displayed explicitly. At saturation $f\cong 2/26$. Since, in an analysis by scattering, the observed intensity is simply proportional to the differential cross section, we can write§§§

$$R=2\frac{d\sigma_+-d\sigma_-}{d\sigma_++d\sigma_-}=2fP_3\frac{d\sigma_1}{d\sigma_0}, \qquad \text{(III-7)}$$

where $d\sigma_1/d\sigma_0$ is the ratio of the two differential cross sections. In order to use this expression it is important to have a value for f that applies to the iron in the region of the scattering. When large solid angles are used, it is also necessary to integrate the cross sections over the angular spreads. In practice the efficiency of the analyzer is often checked with radiation of known polarization.

Since long runs usually are necessary in order to obtain statistically significant results, the changes in the magnetic field are usually made at frequent, periodic intervals in order to minimize the effect of possible electronic drifts. Changes in the counting rate of the detector produced by the changes in the magnetic field are avoided by magnetically shielding the counter and by removing it from the influence of the field by use of a light pipe. If the experimental arrangement is such that changes in magnetic field do affect the counting rate, the effect can be controlled in the beta-gamma coincidence experiments by normalizing the coincidence rate to the gamma-ray singles rate.

§§§ If there is a component of linear polarization in the gamma radiation, the normalization of this expression is changed (Wh 55).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

(b) Measurements on Gamma Rays from Polarized Nuclei (Parity Conserving)

The first experiment dealing with the circular polarization of gamma rays was performed by Clay and Hereford (Cl 52) who attempted to measure the circular polarization of annihilation radiation by detecting Compton electrons ejected in the forward direction from magnetized iron. The effects they observed are now believed to be instrumental in origin (To 56).

Using the apparatus shown in Fig. 32 Wheatley *et al.* (Wh 55) studied circularly polarized gamma rays of nuclear origin in their investigation of the gamma rays emitted from Co^{60} nuclei which had been polarized by the Rose-Gorter method. After adiabatic demagnetization the cryostat containing the source was inserted into the position shown in Fig. 32 (which shows one half of the symmetrical apparatus). It is recalled that the circular polarization is a maximum for gamma rays emitted along the axis of the nuclear polarization, and that the circular polarization changes sign if the axis is reversed. It was convenient therefore to perform the experiment by reversing the nuclear polarization (instead of the electron polarization) which is controlled by the inner magnet M_p (which shares a yoke with the outer magnet M_s). Except for the features associated with the nuclear polarization, especially in the design of the magnet, the polarimeter itself is a forerunner of more recent analyzers using the scattering method. Direct radiation is shielded from the detector, but gamma rays scattered (at angles between 45° and 70°) from the magnetized strip S (Armco iron) are recorded. The magnetization in S is produced by the magnet M_s. The dègree of circular polarization achieved was as high as 75% at the lowest temperature, 0.006°K. The observed fractional change in counting rate upon changing the direction of polarization of the nuclei was at most 3%. In this experiment the circular correlation function has the form (Part I, C)

$$P_3=[A_1P_1(\cos\theta)+A_3P_3(\cos\theta)]/W(\theta),$$

where A_1, A_3, and the coefficients in the directional function $W(\theta)$ are functions of the degree of orientation and so of the temperature. For temperatures in the region $1/T<60$, Wheatley *et al.* obtain excellent agreement with the theoretical correlation function. For lower temperatures the observed deviation may be due to imperfect knowledge of the hyperfine coupling. As we have seen the sign of the effect alone makes possible a determination of the sign of the magnetic moment of the emitting nucleus. In this case the sign of the magnetic moment of the Co^{60} nucleus was found to be positive.

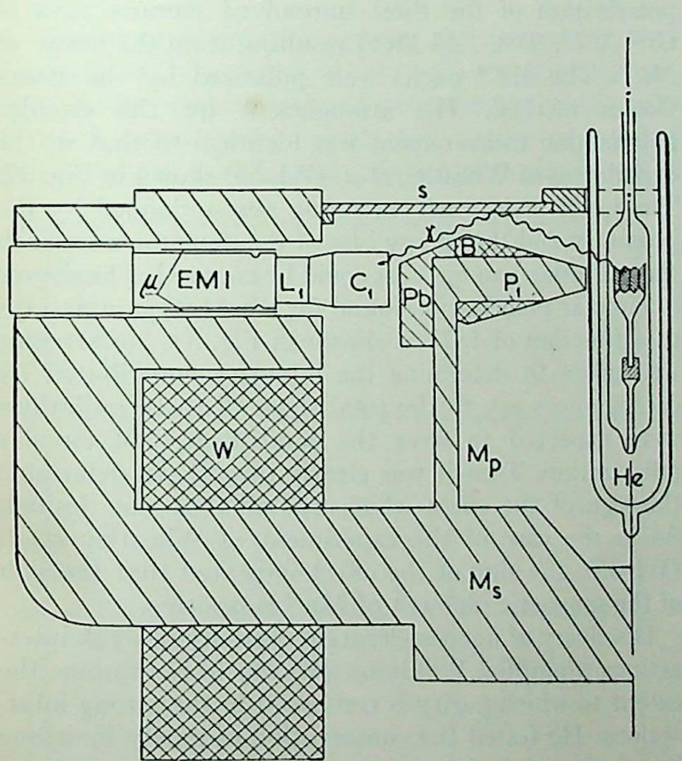


FIG. 32. Diagram of apparatus used by Wheatley *et al.* for producing and measuring circularly polarized gamma rays. A change in the differential Compton cross section is measured by changing the direction of circular polarization of the gamma rays, emitted from polarized nuclei in the cryostat, relative to the direction of magnetization in the scattering iron S. Magnet M_p (with coil B) determines the direction of polarization of the nuclei; Magnet M_s (with coil W) determines the direction of magnetization of S (figure from Wh 55).

Trumpy (Tr 57) using the apparatus shown in Fig. 33 studied the circular polarization of the gamma rays resulting from the capture of polarized neutrons by the following nuclei: S^{32}, Ca^{40}, Ti^{48}, Cr^{53}, Fe^{57}, Ni^{58}, Zn^{64}, and W^{182}. The emitting compound nucleus receives its polarization from the captured neutron. Again, the maximum circular polarization occurs for gamma rays emitted along the axis of polarization. Trumpy selected the transmission method for the analysis of circular polarization. Thermal neutrons emerge from a neutron collimator and pass through a neutron polarizer, consisting of a small iron block mounted in the gap of a magnet in which a field of 14 900 oe is produced. The polarized neutrons impinge on a target and the capture gamma rays, emerging along the directions parallel and antiparallel to the polarization, reach two sodium iodide detectors after being transmitted through the two analyzing magnets. The path through each magnet core is 8 cm long. The heavy shielding of lead and boron carbide reduces the intense background of radiation arising chiefly from neutron capture in the polarizing iron block. Particular gamma rays could be selected for study by means of energy discrimination in the detectors. The experiment was performed by recording the counts in both detectors as a function of the direction of magnetization in the analyzers and the direction of polarization of the neutrons as determined by the polarizing magnet.

Trumpy uses an explicit formula given by Biedenharn *et al.* (Bi 51) for the capture of totally polarized thermal neutrons followed by pure multipole emission. In the correlation $A_1P_1(\cos\theta)$, which measures the degree of

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

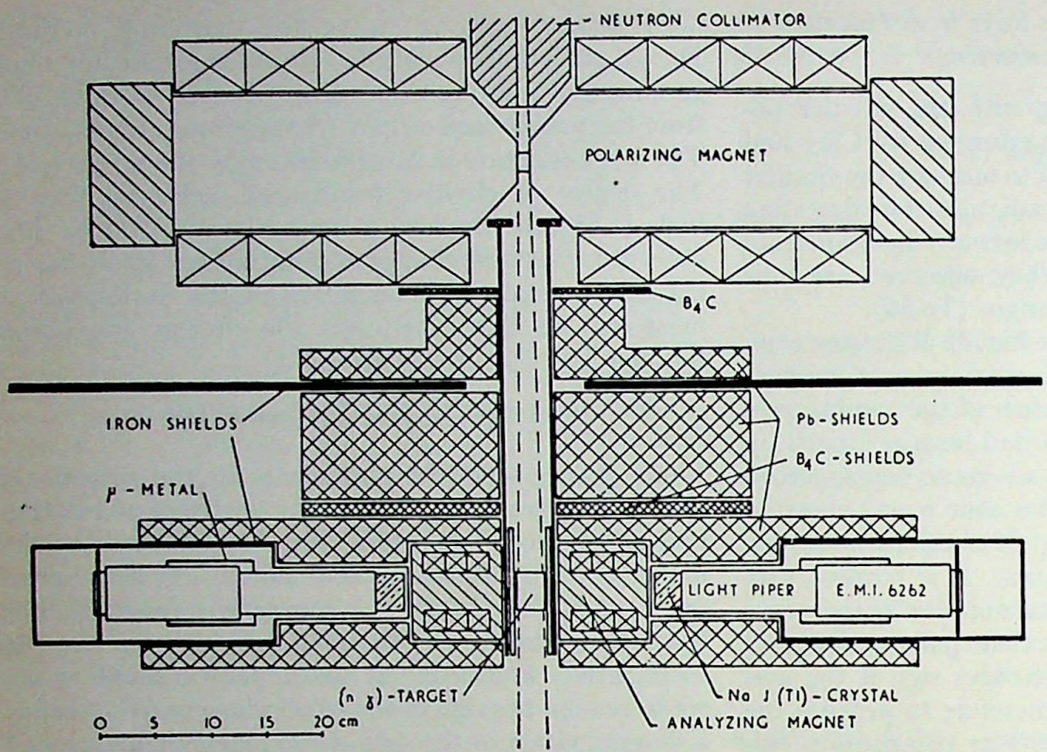


FIG. 33. Top view of experimental arrangement used by Trumpy to detect circular polarization following capture of polarized neutrons. Circular polarization is detected by measurement of the transmission through magnetized iron (figure from Tr 57).

right circular polarization, the coefficient A_1 is given by

$$A_1=2(j-j_1)\frac{L(L+1)+j(j+1)-j_2(j_2+1)}{L(L+1)(2j_1+1)}, \quad \text{(III-8)}$$

where j_1, j and j_2 are the spins of target nucleus, compound nucleus and final nucleus, respectively, and L is the multipole order of the gamma-ray transition. This is equivalent to the result obtained in Part I, C. Since one or more of the angular momenta in (III-8) are often known and $j=j_1\pm\frac{1}{2}$, a measurement of A_1 can yield definite assignments of spin (and parity). For example, if $j_1=0$, $j=\frac{1}{2}$, and $L=1$, then $A_1=1$ or $-\frac{1}{2}$ for $j_2=\frac{1}{2}$ or $\frac{3}{2}$, respectively.

The theoretical value of the circular polarization, given by A_1, is reduced in practice because of the incomplete polarization of the neutron beam (19%). The efficiency of the analyzer was determined in a manner equivalent to (III-5). In the case of S^{33} the spin of the 3.15-Mev level was found to be $\frac{3}{2}$. In Ca^{41} the spin of the 1.98-Mev level also was determined as $\frac{3}{2}$. The measurements on the 6.41- and 6.75-Mev gamma rays from Ti^{49} were consistent with spins of $\frac{1}{2}$ and $\frac{3}{2}$ for the 1.70- and 1.35-Mev states, respectively. Trumpy's results for Cr^{54} in conjunction with his angular correlation work (Tr 57a) indicated a 1(1)2(2)0 decay scheme for the 9.72-, 0.84- and 0-Mev levels. Also both the 9.72- and 8.88-Mev gamma rays are $E1$. Contrary to shell-model predictions, the results on Fe^{57} indicated that its ground state is $\frac{1}{2}^-$. A spin of $\frac{3}{2}$ was confirmed for the ground state of Ni^{59}, and either $\frac{3}{2}$ or $\frac{5}{2}$ for the ground state of Zn^{65}. The spins and multipolarity involved in the decay of the 6.18-Mev state in W^{183} were known from other work. Consequently the measurement on this nucleus constituted a verification of the direction and order of magnitude of the expected circular polarization.

In another part of their work, discussed in Part III, A(f), Huiskamp *et al.* (Hu 57) measured the circular polarization of the three unresolved gamma rays of Cr^{52} (0.73, 0.94, 1.43 Mev) resulting from the decay of Mn^{52}. The Mn^{52} nuclei were polarized by the Rose-Gorter method. The arrangement for the circular polarization measurement was identical to that in the experiment of Wheatley *et al.* (Wh 55) shown in Fig. 32. The range of forward scattering angles defined by the geometry and the energy bias of the detector were such that all three gamma rays could be counted. The degree of circular polarization found for the Mn^{52} gamma rays as a function of $1/T^*$ is shown in Fig. 34. No attempt was made to determine the intensity contributed by each gamma ray to the total since all three radiations were expected to have the same degree of circular polarization. Thus it was clearly possible to determine the sign of the effect which was all that was desired. As in the case of the experiment of Wheatley *et al.* (Wh 55), the sign of the effect indicated that the sign of the magnetic moment of Mn^{52} is positive.

Discovery of nonconservation of parity in weak interactions prompted Wilkinson (Wi 58) to re-examine the extent to which parity is conserved in the strong interactions. He tested the conservation of parity in strong interactions by three classes of experiments. One of these involves a search for circular polarization of gamma rays from a nuclear reaction in the initial state of which all particles are unpolarized. Wilkinson looked for circular polarization of the 2.14- and 7.12-Mev gamma rays from the $B^{11}(p,p')B^{11*}(\gamma)B^{11}$ and $F^{19}(p,\alpha)O^{16*}(\gamma)O^{16}$ reactions, respectively. He used

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

transmission through magnetized iron as the method of analysis for circular polarization. Since very small effects, if any, were expected, considerable precautions were taken to eliminate systematic errors resulting, for instance, from the influence of magnet reversals on the photomultipliers or from electronic drifts. The scintillating crystal was mounted at the end of a light pipe either two or four feet long. The two-foot pipe was used with the $F^{19}(p,\alpha)O^{16*}$ reaction for which greater resolution was needed. The photomultiplier was surrounded not only with a mu-metal shield but also with two coaxial iron pipes. To prevent the degraded radiation emerging from the magnet from masking the effect under study, the bias of the electronic apparatus was set to accept only the very high-energy end of the spectrum. Since the search was for such small effects, over five-hundred runs were made on each reaction. The results of the runs were carefully averaged in different ways to eliminate the possible influence of systematic errors. Average values of $N+/N-=1.00025 \pm 0.00026$ and 1.00003 ± 0.00024 were found for the boron and fluorine reactions, respectively, where $N+/N-$ is the ratio of the counting rate with the magnetic field in one direction to that in the other. Thus, within the aforementioned limits, there is no circular polarization and no nonconservation of parity.

(c) *Beta-Gamma Circular Polarization Correlation*

The measurement of circular polarization in beta-gamma decay was one of the early experiments performed after the discovery of nonconservation of parity in weak interactions (Le 56) and the revival of the two-component theory of the neutrino (Le 57). The

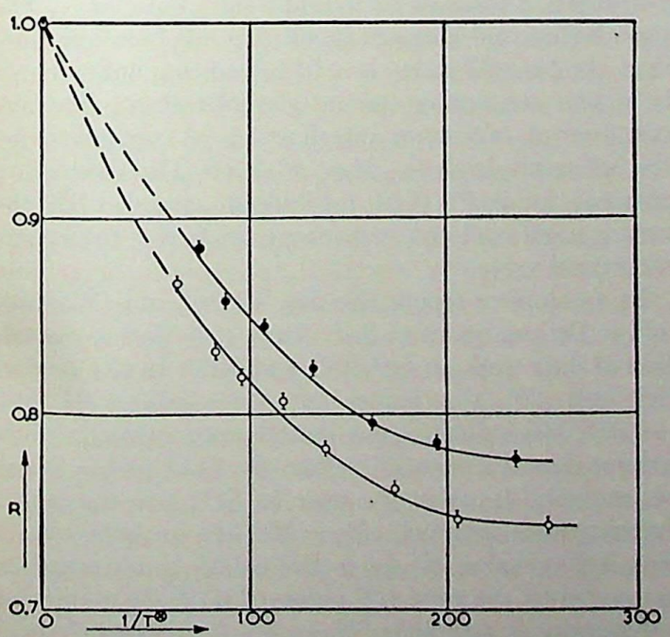


FIG. 34. Circular polarization of gamma rays emitted from a polarized Mn^{52} source. Ordinate R is the counting rate normalized to unity at $T=1°K$. Open circles correspond to the polarizing field and the induction in the scattering iron being parallel, while closed circles correspond to their being antiparallel (figure from Hu 57).

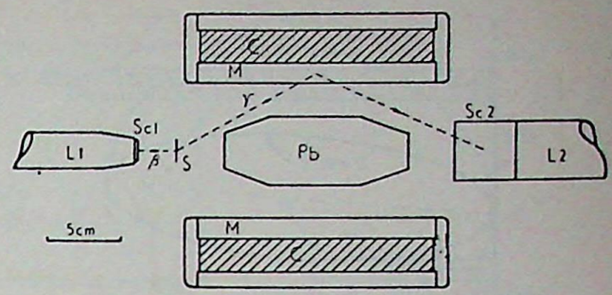


FIG. 35. Experimental arrangement used by Schopper for the beta-gamma circular-polarization correlation. Sc 1 is the beta counter using a plastic scintillator; Sc 2 is the gamma counter with a NaI(Tl) crystal; L1 and L2 are light guides; S is the source; and M is a cylindrical magnet with magnetizing coils C_1 (figure from Sc 57).

experiments have since been carried forward and have provided valuable information on the nature of beta decay. The aim of the measurements is to determine the asymmetry coefficient $A_1=-(v/c)A$ in the circular polarization correlation (Part I, C),

$$P_3=(v/c)A\cos\theta, \qquad \text{(III-9)}$$

which gives the degree of right circular polarization as a function of the angle between the beta particle and the gamma ray. Both the scattering method and the transmission method have been used to good advantage in this work, the polarizations being obtained by (III-5) and (III-7) or their equivalent. The beta-gamma circular polarization measurement provides the same information as does a knowledge of the electron distribution from polarized nuclei. The former technique has the advantage, however, of being feasible for a large number of different nuclei.

Schopper (Sc 57) first performed such a correlation experiment in the decay of Co^{60} and of Na^{22}. These nuclei have allowed‖ ‖ ‖ β transitions with $\Delta J=1$ followed by pure multipole radiation. The experimental arrangement (Fig. 35) is designed to study the circular polarization of the gamma rays emitted at an angle close to 180° measured with respect to the direction of the beta particles. The average forward scattering angle was 55°, which is about optimum for circularly polarized gamma rays of the energies used in the experiment (~1.25 Mev). Electronic discrimination in the circuits was provided in order to avoid counting backscattered gamma rays, annihilation radiation, and stray beta particles. Although light pipes were used in the detectors, possible variations in the coincidence rate produced by the changes in the magnetic field were eliminated by using the ratio of the coincidence rate to the product of the singles rates.

In the decay of Co^{60} two gamma rays are emitted in a 4–2–0 cascade, but the gamma rays can be treated as one since in this special case the polarization of the second gamma ray is the same as that of the first

‖ ‖ ‖ All of the beta transitions discussed in this section are allowed unless otherwise specified.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

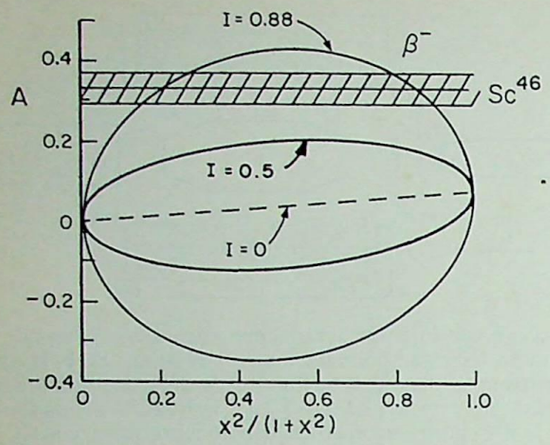


FIG. 36. Asymmetry coefficient in the beta-gamma circular polarization for a $4^+(\beta)4^+(\gamma)2^+(\gamma)0^+$ transition. The experimental value for Sc^{46} is indicated. Theoretical curves for $I=I_{max}=0.88$, $I=0.5$ and $I=0$ are shown; I is the interference term that enters into A. If $I>0$, the upper branch of the theoretical curve refers to the choice of $x=a^{\frac{1}{2}}M_{GT}/M_F>0$ and the lower branch to the case $x<0$. If $I<0$ the opposite assignment is true. The figure shows that a large value for I is required to fit the experimental data (figure from Bo 58).

(To 53). Schopper found $A=+0.39\pm0.08$ and -0.41 ± 0.07 for Na^{22} and Co^{60}, respectively. These results can be compared to theoretical values of $+\frac{1}{3}$ and $-\frac{1}{3}$ respectively. The result on Co^{60} is in accord with the experiment of Wu *et al.* (Wu 57) which measured the electron distribution from polarized Co^{60} nuclei. The opposite signs for Co^{60} and Na^{22} indicate opposite circular polarizations and show that the antineutrino emitted in negatron decay and the neutrino emitted in positron decay have opposite helicity. In a note added in proof in his paper Schopper reports a result leading to a value of $A=-0.07\pm0.05$ for Na^{24}.

Boehm and Wapstra (Bo 57, Bo 58) using a very similar experimental arrangement have studied the circular-polarization correlations in many beta-gamma decays. Their first experiments were performed on Co^{60}, Au^{198} and Hg^{203}, the latter two of which have first-forbidden beta transitions to the 0.411- and 0.279-Mev states in Hg^{198} and Tl^{203}, respectively. A refined expression for the efficiency of the analyzer was calculated by Alder,

$$\epsilon=2.90k(1+0.13k)/(1+0.36k+0.09k^2),$$

where ϵ would be the experimental effect (in percent) for a completely circularly polarized gamma ray of energy k (in units of m_0c^2). Measurements on the circularly polarized bremsstrahlung created by the beta-particles from P^{32} and Tm^{170} were made partly to check this formula and partly to establish the v/c law for the polarization of electrons emitted in beta decay. The experimental results found for P^{32} and Tm^{170} were consistent with both the formula and the v/c law. In the circular-polarization measurement performed on Co^{60} a value of $A=-0.41\pm0.08$ was found, in agreement with Schopper's value. The value of $A=0.52\pm0.16$ that was found for Au^{198} is consistent with a spin of 2 and rules out a spin of 3 for the ground state of Au^{198}. With the value of $A=-0.06\pm0.22$ for Hg^{203} (Bo 58) it was not possible to exclude a spin of $\frac{5}{2}$ for the ground state of Hg^{203} but the measurement is consistent with a spin of $\frac{3}{2}$.

Boehm and Wapstra studied further the above J-J transiton in Au^{198} along with those in Co^{58} (Bo 57a, Bo 58) and Sc^{46} (Bo 57b, Bo 58). The J-J transitions are of great interest because of the mixed nature of the beta transition. Thus the decay scheme of Sc^{46} is $4^+(\beta)4^+(2)2^+(2)0$ and differs from the Co^{60} case only in that $\Delta J=0$ instead of 1. For this decay scheme, Boehm and Wapstra write the asymmetry coefficient in compact form (Part I, C) as

$$A=(0.0834x^2+0.745a^{\frac{1}{2}}Ix)/(1+x^2), \quad \text{(III-10)}$$

where $x=a^{\frac{1}{2}}M_{GT}/M_F$, $a=(C_{GT}/C_F)^2=1.3$ (Ko 56) and I is the interference factor whose numerator is equal to S_1 for $LL'=01$ in Table X. If $I=0$, A varies from 0 to 0.083, as x goes from 0 to ∞. If $Ix>0$, then $A>0$, but if $Ix<0$, then $A<0$ except when $|Ix|<0.098x^2$. Finally, if the two-component neutrino theory with invariance under time reversal is valid, the maximum value of $|I|$ is given by $a^{-\frac{1}{2}}=0.88$. These properties are illustrated in Fig. 36 which gives curves for three values of $|I|$, namely 0, 0.5 and the maximum 0.88. The experimental result for Sc^{46}, $A=0.33\pm0.04$, is shown on the graph. The experimental value indicates a large value of $|I|$. Boehm and Wapstra give $|I|\geq0.5$ with a statistical uncertainty of 1%. Since the dominant term¶¶¶ in I is $\mathrm{Re}(C_SC_T'^*+C_S'C_T^*-C_VC_A'^*-C_V'C_A^*)$, this large value of $|I|$ rules out a pure V, T or a pure S, A interaction in this beta decay. If the maximum interference is assumed, a value of 2.2 is found for $|M_{GT}/M_F|$. For Co^{58}, Boehm and Wapstra found $A=-0.14\pm0.07$, and since the pure GT value is -0.083, no conclusion could be drawn concerning the interference term. The assumption of maximum interference, as suggested by the Sc^{46} result, leads to $|M_{GT}/M_F|>8$. The value of A obtained for Au^{198} (first forbidden) agreed with the earlier result and was consistent with the maximum theoretical value.

In a complete report (Bo 58), Boehm and Wapstra add to the treatment of Sc^{46}, Au^{198}, and Co^{58} a discussion of their work on Sc^{44}, V^{48}, and Na^{24}. In the case of Sc^{44} and V^{48}, the values of $A=-0.02\pm0.04$ and $+0.06\pm0.05$ indicate some interference since they are greater than the pure GT values of -0.17 and -0.083, respectively. If, as in the case of Sc^{46}, the maximum interference is assumed, $|M_{GT}/M_F|\cong5$ for both nuclei. For Na^{24} a value of $A=+0.07\pm0.04$ is obtained as compared to the pure GT value of 0.08. If maximum interference is assumed, $|M_{GT}/M_F|>25$. Interference between the GT and F interactions was also found by Boehm (Bo 58a) in the decay of Mn^{52}. In this example, $A=-0.16\pm0.05$, and the pure GT value is -0.056.

¶¶¶ The other term in I has a maximum value $\cong0.12$.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by eGangotri Foundation

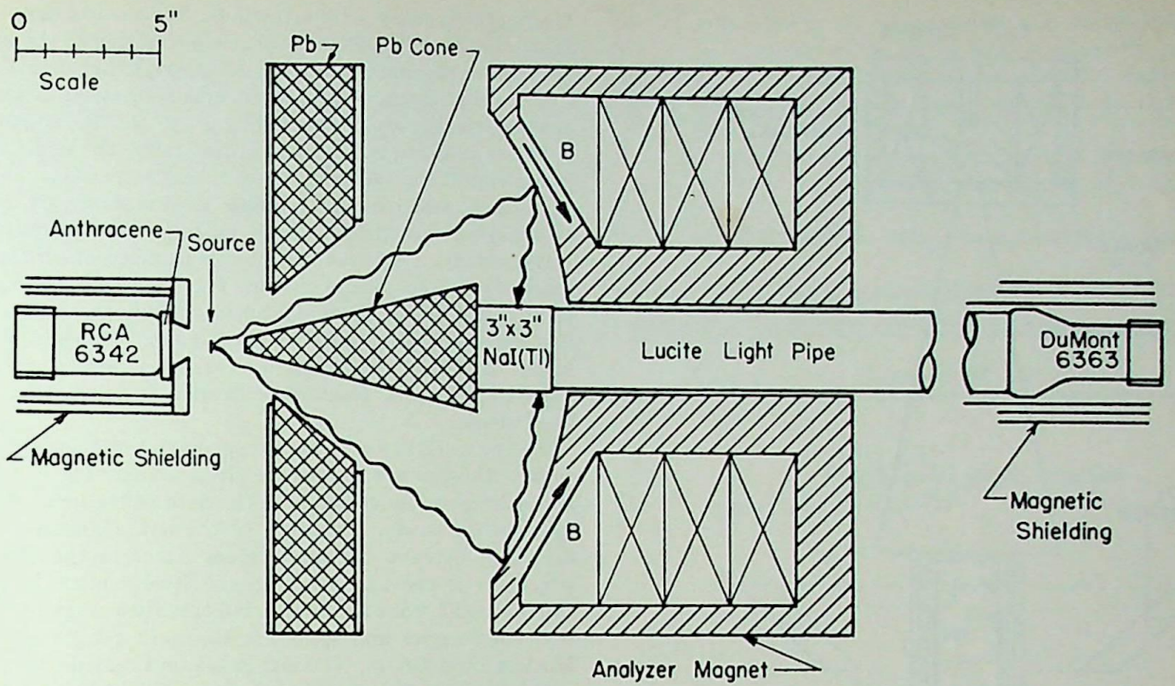


FIG. 37. Experimental arrangement of Steffen utilizing the method in which the electron polarization is perpendicular to the circular polarization (figure from St 58a).

Among the other early investigators in this field were Lundby *et al.* (Lu 57) who also made measurements on Co^{60}. Since it was felt the energy discrimination would be improved, they used the method of transmission through magnetized iron instead of the forward-scattering technique. They found a value of $A=-0.32\pm0.07$. In a note added in proof a value of $A=+0.34\pm0.14$ for Na^{22} also is given.

DeBrunner and Kündig (De 57) using the method of scattering also measured the anisotropy coefficient for Co^{60} and found $A=-0.344\pm0.09$ in good agreement with other values and with the theoretical value of $-\frac{1}{3}$. Shortly thereafter Berthier *et al.* (Be 57a) using the same equipment studied Au^{198}, the decay of which is first forbidden. It was found that $A=0.34\pm0.09$ which is somewhat smaller than the value above (Bo 58).

Again using the forward scattering from magnetized iron, Appel and Schopper (Ap 57) measured the circular-polarization correlations in Co^{60}, Zr^{95}, and Sb^{124}. The measurement on Co^{60} was primarily to test the reliability of the apparatus, but they give a value of $A=-0.35\pm0.05$ for this nucleus. In the case of the decay of Zr^{95}, the measurements were made on the unresolved 0.722- and 0.754-Mev radiations from Nb^{95}. By comparing the experimental value of $A=-0.46\pm0.09$ with possible theoretical values in the manner discussed in the foregoing (Al 57), Appel and Schopper conclude that there is probably present an ST or VA interference term. The measurements made with Sb^{124}, which has a first-forbidden transition to the 0.605-Mev state in Te^{124}, yielded a value of $A=-0.13\pm0.06$. This is smaller in absolute value than the value of $A=-\frac{1}{3}$ calculated (Al 57) on the basis of a simple once-forbidden transition.

In an effort to determine whether or not parity breakdown is a maximum, Appel *et al.* (Ap 58) remeasured more accurately the beta-gamma circular-polarization correlations in Co^{60} and Na^{22}. The measurement of the polarization correlation seems favorable for this purpose since all the necessary corrections can be calculated accurately and systematic errors can be kept small. They found $A=-0.340\pm0.035$ and $A=+0.295\pm0.054$ for Co^{60} and Na^{22}, respectively. These values compare favorably with the corresponding theoretical values of $-\frac{1}{3}$ and $+\frac{1}{3}$ for the case when nonconservation of parity and noninvariance under charge conjugation are maximum.

Steffen (St 58) also studied the circular polarization of the gamma rays from Co^{60} and Sc^{46}. He verified the dependence of the degree of circular polarization on v/c [Eq. (III-9)] and that the angular distribution varies as $\cos\theta$. For Co^{60} and Sc^{46}, values of $A=-0.34\pm0.02$ and $A=+0.24\pm0.02$ were found. Although the Sc^{46} value is smaller than that given by Bo 58, it verifies the presence of a large interference between GT and F interaction. If maximum interference between V and A interactions is assumed, $|M_{GT}/M_F|\simeq4.0$. With these nuclei Steffen (St 85a) has also used the method (Be 57) in which the electron polarization is perpendicular to the circular polarization. His experimental arrangement is shown in Fig. 37.

In a beautiful experiment by Goldhaber *et al.* (Go 58) on the 80-kev isomer of Eu^{152}, it was shown that the neutrino is left handed (negative helicity). Eu^{152} decays by electron capture to the 961-kev state in Sm^{152}. The experiment was performed by measuring the circular polarization of the 961-kev gamma rays which are subsequently resonant scattered from Sm^{152}. The ex-

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

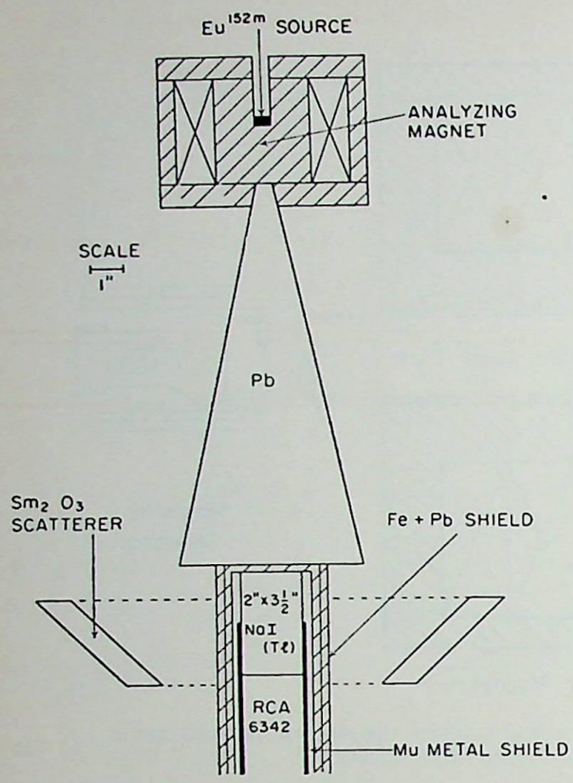


FIG. 38. Experimental arrangement for measurement of circular polarization of resonant scattered gamma rays. This arrangement was used by Goldhaber *et al.* to determine helicity of neutrinos (figure from Go 58).

perimental arrangement is shown in Fig. 38. Through conservation of energy and momentum the requirement of resonant scattering selects those gamma rays which are emitted opposite to the neutrino emitted in the electron capture. The magnetic analyzer determines the circular polarization (helicity) of the gamma rays. The helicity of the neutrino is then determined since a gamma ray emitted in the direction opposite to the neutrino will have the same helicity by conservation of angular momentum. The measurement gave negative helicity for the gamma ray and thus also for the neutrino.

E. Tabulation of Polarization Experiments on Nuclear Gamma Rays

The methods and results of the experiments discussed in the text are summarized in Table XIII. An explanation of the notation used and how the entries were made in the various columns of the table is given below.

Column I. Nucleus.—In all cases the nucleus given in this column is the nucleus emitting the gamma ray whose polarization is being measured. The experiments are listed in order of increasing mass number A.

Column II. Source of reaction.—In this column is given a level diagram of the radioactive decay or nuclear reaction used in each experiment. Frequently one diagram is applicable to several experiments and so is not repeated. The *level diagrams given in the table are not complete.* Only those levels which directly feed the gamma-ray transitions under study and those between the contributing levels and the ground state are shown in the diagram. Secondly, the level diagrams show only those transitions in the daughter or residual nucleus which give rise to the polarizations which were measured. Thirdly, only those beta transitions which lead directly to the gamma ray or cascade of gamma rays under study are given by solid arrows. Dashed arrows denote the fact that other beta transitions *may* be contributing to the intensity of the gamma ray under study but not through the particular cascade under study. Decay by positron emission, electron emission, or electron capture are designated by β^+, β^-, or EC, respectively.

In the case of the reactions in this table, the neutron, proton, or alpha particle is denoted by n, p, or α, respectively. In the case of resonant reactions, the resonant level is shown by a solid line. Nonresonant reactions show a dashed line, whereas in thick-target reactions a bracket is shown in place of all the "levels" reached in the compound nucleus. The bombarding energy is given as a dashed line in the initial nucleus. Neutron-capture reactions are shown with the "state" of the compound nucleus at the same height as that of the initial nucleus since thermal capture is involved in all cases. Coulomb excitation is denoted by a double-lined arrow.

Energy levels below 10 Mev are given to three significant figures, some of those above 10 Mev are given to four. The energies of the levels are given immediately to the right of the level, whereas the spins, when known, are given to the left. Gamma-ray energies are given between the level values either to the right or left, whichever is convenient. When the level scheme is sufficiently complicated, arrows are used to indicate the energy of the gamma ray. All energies and spins are the most recent given by the Nuclear Data Group, National Research Council (Mc 58a).

Column III. Polarization technique.—The letters given in this column denote the basic polarization technique used in each of the experiments. They are as follows:

C = Compton polarimeter detecting linear polarization.
C_f = Compton magnetic polarimeter detecting circular polarization by forward scattering.
C_t = Compton magnetic polarimeter detecting circular polarization by transmission.
D = Photodisintegration of the deuteron, detecting linear polarization by use of nuclear emulsions soaked in D_2O.

Column IV. Experimental conditions.—Three essential experimental conditions are given in this column. In order from top to bottom, as given with each experiment, they are as follows. (1) The physical entity which establishes the direction of quantization in the polarization correlation. Thus, if the direction is established by the propagation vector of an alpha particle, the entry α is made; if by a magnetic field, the entry "mag. field" is made, etc. In cases where the direction of propagation of a gamma ray in a cascade establishes the direction, the gamma ray is specified by energy when it is known. If there has been no clear discrimination between which gamma ray establishes the direction and which is accepted by the polarimeter, only the symbol γ is placed in this position. It should be noted in cases where Bleaney's method of alignment is used, that the words "crystal electric field" have been used primarily for convenience and do not in general mean that this field direction is the alignment direction. In general the alignment direction is established as the result of the coupling between the crystal electric field and the atomic orbital motion which in turn is coupled to the spins. Thus different alignment directions may occur depending on the crystal used. (2) Below entry (1) is given the angle between the axis of quantization and the axis of the polarimeter. When a complete direction-polarization correlation function is determined, the range of angles is indicated by an arrow between the extremes of the angular range. (3) Below entry (2) is specified the gamma ray whose polarization is being measured. When there is more than one gamma ray being analyzed and energy discrimination is made between each, a semicolon is placed between them. When the gamma rays are not resolved, a comma is placed between them and a (u) is inserted at the end. When there is discrimination available, but it is not specified which of the gamma rays is being analyzed, an "or" is placed between the gamma-ray entries.

Column V. Summary of results.—In the case of linear polarization measurements, the gamma rays studied are usually listed

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

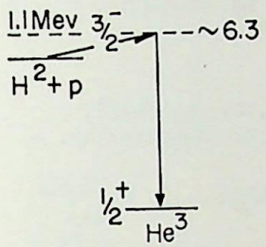
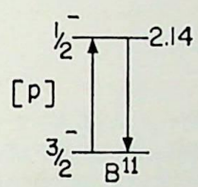
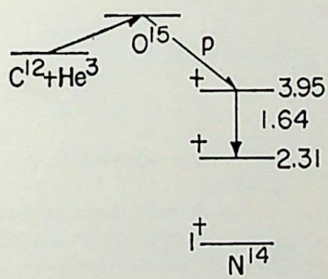
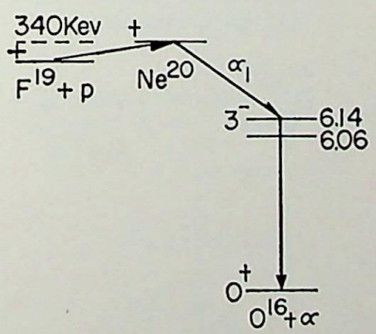
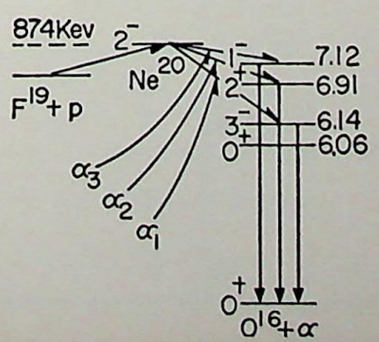
first in vertical order with their multipolarities. The multipolarity is given even though additional evidence outside of the experiment listed usually is required to deduce it. In the case of the circular polarization measurements the anisotropy coefficient A is given for each gamma ray listed. In two instances N_+/N_- is used to indicate the ratio of counting rates before and after reversing the magnetic field.

The parity of a state is listed when it is of special interest, usually as designated by the author. The parity is determined, of course, only relative to other assumed or known parities. Other information such as spins, the sign of a magnetic moment, etc., is frequently given also when it is considered a significant part of the author's conclusions. As a general rule, the chief conclusion of the author is the one given.

A colon is used to connect an item given in the column with the results of the measurements made on it, e.g., "cascade: $4(E2)2(E2)0$."

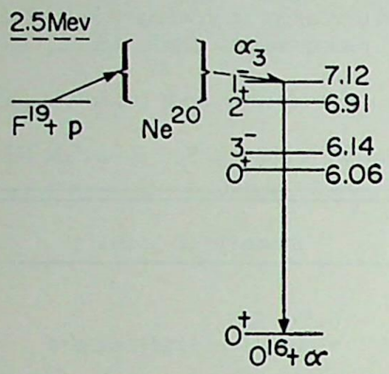
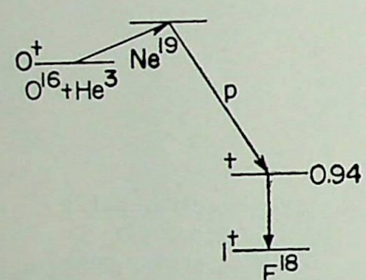
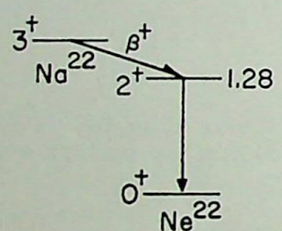
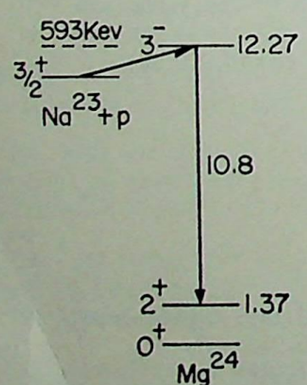
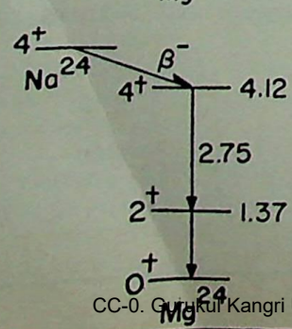
A reference is made to other work (usually a measurement of

TABLE XIII. Summary of experiments. An explanation of the various entries is given in Sec. E of Part III.

Nucleus	Source or Reaction	Polarization Technique	Experimental Conditions	Summary of Results	Reference
He^3		D	p 90° ~6.3γ	6.3γ: E1, ~100% polarized; agrees with Fo49	Wi52
B^{11}		Ct	p 0° 2.14γ	2.14γ: no circ. pol.; N_+/N_- = 1.00025 ± 0.00026, parity conserved	Wi58
N^{14}		C	He^3 90° 1.64γ	1.64γ: M1 or E2, 3.95 state parity: +	Li58
O^{16}		C	α 118° 6.14γ	6.14γ: E3; Ne^{20} exc.state and F^{19} gnd.state parities: + (Ar50, Ba50)	Fr52
		D	p 90° 6.14γ;6.91γ; 7.12γ	6.14γ: unpolarized, 6.91γ: E2, 7.12γ: E1; agrees with Se52 (Sa52)	Fa53

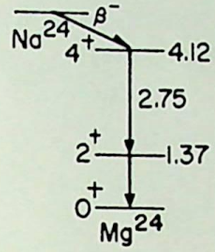
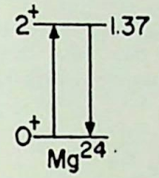
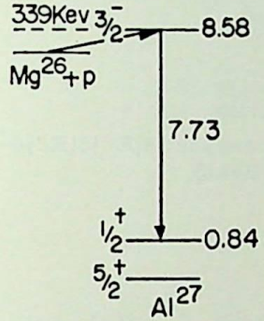
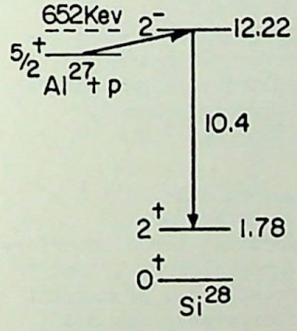
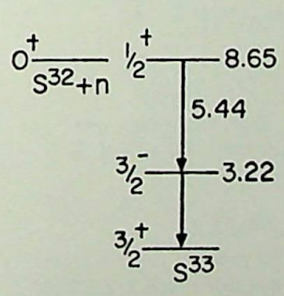
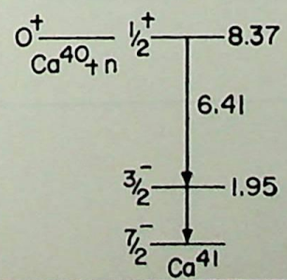
CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

TABLE XIII (*continued*).

Nucleus	Source or Reaction	Polarization Technique	Experimental Conditions	Summary of Results	Reference
O^{16}		C_t	p 0° 7.12γ	7.12γ: no circ. pol. $N_+/N_- = 1.00003 \pm 0.00024$ parity conserved	Wi58
F^{18}		C	He^3 90° 0.94γ	0.94γ: M1 or E2, 0.94 state parity: +	Li58
Ne^{22}		C_f	β 152° 1.28γ	1.28γ: A = -0.39 ± 0.08; additional proof of nonconservation of parity.	Sc57
				1.28γ: A = -0.340 ± 0.035; accurate indication that nonconservation of parity and noninvariance under charge conjugation are maximum	Ap58
Mg^{24}		D	p 90° 10.8γ	10.8γ: E1, Na^{23} gnd.state parity: + (Gr55)	Hu56
		C	γ 100° 1.37γ or 2.75γ	1.37γ: E2, 2.75γ: E2; agrees with B152	Es56

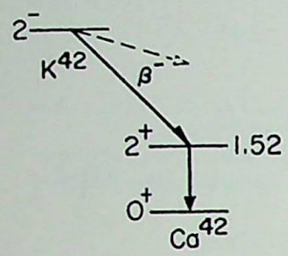
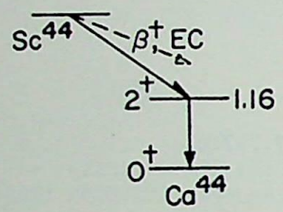
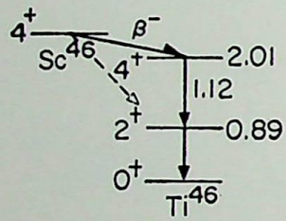
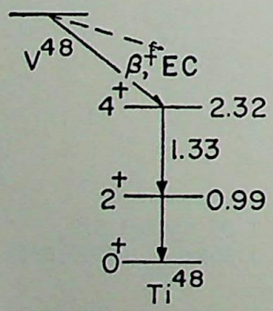
CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

TABLE XIII (*continued*).

Nucleus	Source or Reaction	Polarization Technique	Experimental Conditions	Summary of Results	Reference
Mg^{24}		C_f	β 152° 1.37γ,2.75γ(u)	1.37γ, 2.75γ(u): A = -0.068 ± 0.047	Sc57 Ap57
		C_f	β 159° 1.37γ,2.75γ(u)	1.37γ, 2.75γ(u): A = +0.07 ± 0.04; if max.interference assumed, $\lvert M_{GT}/M_F \rvert > 25$	Bo58
		C	p 90° 1.37γ	1.37γ: E2	Li58
Al^{27}		D	p 90° 7.73γ	7.73γ:E1, 0.84 state parity: + agrees with Da53 (Ru56)	Hu56
Si^{28}		D	p 90° 10.4γ	10.4γ: E1, Al^{27} gnd.state parity: - (Ru54)	Hu56
S^{33}		C_t	n polarization 0°,180° 5.44γ	3.22 state: 3/2, 3.22γ: E1 (Ho53,Br54)	Tr57
Ca^{41}		C_t	n polarization 0°,180° 6.42γ	1.95 state: 3/2, 1.95γ: E2 (Ho53,Br54)	Tr57

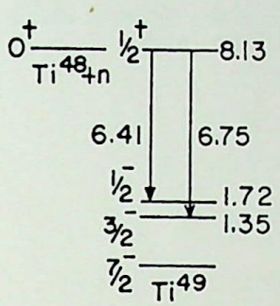
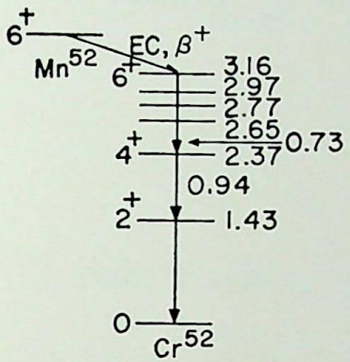
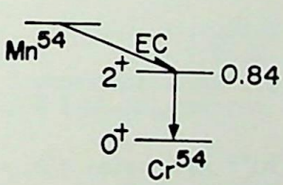
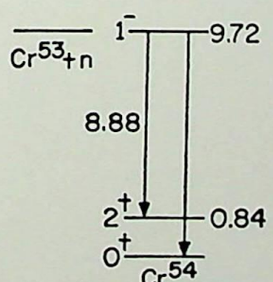
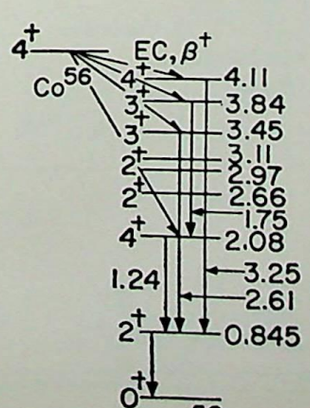
CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

TABLE XIII (*continued*).

Nucleus	Source or Reaction	Polarization Technique	Experimental Conditions	Summary of Results	Reference
Ca^{42}		C	β 90° 1.52γ	1.52γ: unpolarized (Be50b, St51)	Ha53a
Ca^{44}		C_f	β 159° 1.16γ	1.16γ: A = -0.02 ± 0.04; if max.interference assumed, $\|M_{GT}/M_F\| \cong 5$	Bo58
Ti^{46}		C	γ 90° → 180° 0.89γ,1.12γ(u)	0.89γ: E2, 1.12γ: E2. cascade: 4(E2)2(E2)0 (Br48)	Me50
		C_f	β 159° 0.89γ,1.12γ(u)	0.89γ, 1.12γ(u): A = 0.33 ± 0.04; large interference between GT and F couplings; if max.interference assumed, $\|M_{GT}/M_F\| \cong 2.2$	Bo57b Bo58
		C_f	β ~155° 0.89γ,1.12γ(u)	0.89γ, 1.12γ(u): A = 0.24 ± 0.02; verifies large interference between GT and F couplings; if max.interference assumed, $\|M_{GT}/M_F\| \cong 4$	St58
Ti^{48}		C_f	β 159° 0.99γ,1.33γ(u)	0.99γ, 1.33γ(u): A = 0.06 ± 0.05; if max.interference assumed, $\|M_{GT}/M_F\| \cong 5$	Bo58
		C	1.33γ; 0.99γ 105° 0.99γ; 1.33γ	0.99γ:E2 133γ:E2 cascade: 4(E2)2(E2)0	Co56a

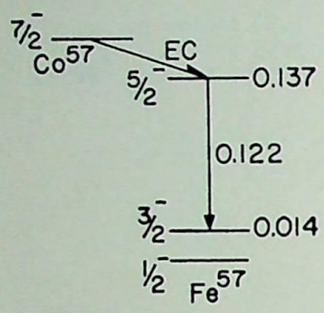
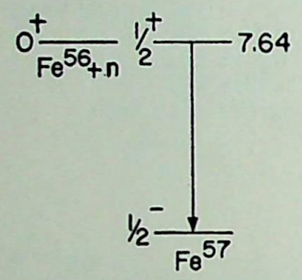
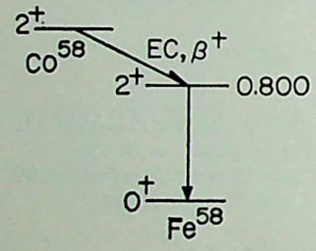
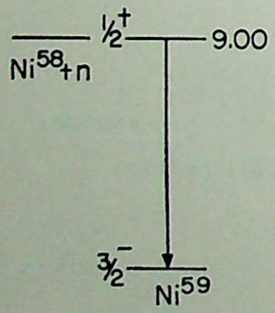
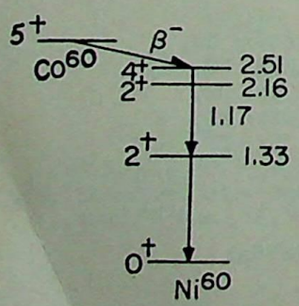
CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

TABLE XIII (*continued*).

Nucleus	Source or Reaction	Polarization Technique	Experimental Conditions	Summary of Results	Reference
Ti^{49}		C_t	n polarization 0°,180° 6.41γ;6.75γ	1.72 state: ½, 1.35 state: 3/2, 1.35γ: E2 (Ho53,Br54)	Tr57
Cr^{52}		C	mag.field 90° 0.73γ;0.94γ; 1.43γ	0.73γ: E2, 0.94γ: E2, 1.43γ: E2; agrees with Ke54	Hu57
		C_f	mag.field 0°,180° 0.73γ,0.94γ, 1.43γ(u)	Mn52 mag. mom:+	Hu57
		C_f	β 159° 0.73γ,0.94γ, 1.43γ(u)	0.73γ, 0.94γ, 1.46γ(u): A = -0.16 ± 0.05; interference between GT and F couplings	Bo58a
Cr^{54}		C	mag.field 90° 0.84γ	0.84γ: E2	Bi54
		C_t	n polarization 0°,180° 8.88γ;9.72γ	8.88γ: E1, 9.72γ: E1; 8.88 - 0.84 cascade: 1(E1)2(E2)0	Tr57
Fe^{56}		C	mag.field 90° 0.845γ;1.24γ; 1.75γ:2.61γ: 3.25γ	0.845γ: E2, 1.24γ: E2, 1.75γ: M1, 2.61γ: M1, 3.25γ: E2; 0.845,2.08,3.45,3.84, and 4.10 levels: 2+,4+, 3+,3+, and 4+ resp. (Sa55a)	Di58
		C	1.24γ 90° 0.845γ	0.845γ:E2 (Hu55)	Wo55

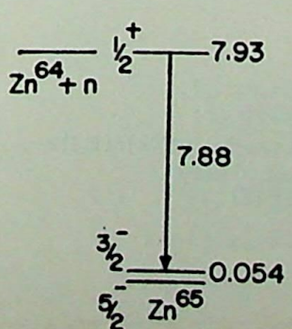
CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

TABLE XIII (*continued*).

Nucleus	Source or Reaction	Polarization Technique	Experimental Conditions	Summary of Results	Reference
Fe^{57}		C	crystal electric field 90° 0.122γ	0.122γ: E2/M1 = +0.19 ± 0.02 (A154)	Bi55a
		C_t	n polarization 0°,180° 7.64γ	Fe^{57} gnd.state: $\frac{1}{2}$ (B153)	Tr57
Fe^{58}		C	crystal electric field 90° 0.800γ	0.800γ: E2 (Da52)	Bi52
		C_f	β 159° 0.800γ	0.800γ: A = -0.14 ± 0.07; if max.interference assumed, $\lvert M_{GT}/M_F \rvert \cong 2.8$	Bo57a Bo58
Ni^{59}		C_t	n polarization 0°,180° 9.00γ	Ni^{59} gnd.state: 3/2; confirms Pr54	Tr57
Ni^{60}		C	γ 90° → 180° 1.17γ,1.33γ(u)	1.17γ: E2, 1.33γ: E2; 1.17 - 1.33 cascade: 4(E2)2(E2)0 (Br48)	Me50
			γ 90° → 180° 1.17γ,1.33γ(u)	1.17γ: E2, 1.33γ: E2; agrees with Me50	Wi50

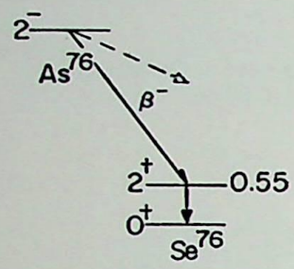
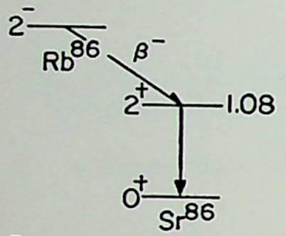
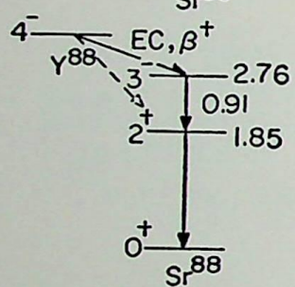
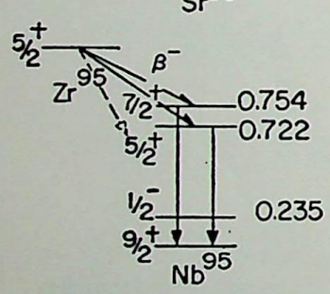
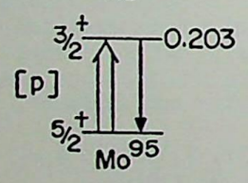
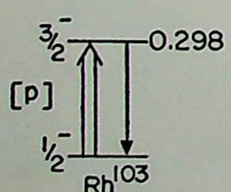
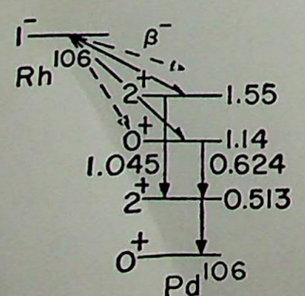
CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

TABLE XIII (*continued*).

Nucleus	Source or Reaction	Polarization Technique	Experimental Conditions	Summary of Results	Reference
Ni^{60}		C	γ $90^\circ \rightarrow 180^\circ$ 1.17γ,1.33γ(u)	1.17γ: E2, 1.33γ: E2; agrees with Me50 and Wi50	Kl52
		C	crystal electric field 90° 1.17γ,1.33γ(u)	1.17γ: E2, 1.33γ: E2; agrees with Me50	Bi52
		C	γ 100° 1.17γ or 1.33γ	1.17γ: E2, 1.33γ: E2; agrees with Me50	Es56
		C_f	mag.field 151° 1.17γ,1.33γ(u)	Co^{60} mag. mom: +	Wh55
		C_f	β 152° 1.17γ,1.33γ(u)	1.17γ, 1.33γ(u): A = -0.41 ± 0.07; ν in β^+ decay in opposite screw sense from ν^- in β^- decay	Sc57
		C_f	β 159° 1.17γ,1.33γ(u)	1.17γ, 1.33γ(u): A = -0.41 ± 0.08	Bo57 Bo58
		C_t	β 180° 1.17γ,1.33γ(u)	1.17γ, 1.33γ(u): A = -0.32 ± 0.07	Lu57
		C_f	β 153° 1.17γ,1.33γ(u)	1.17γ, 1.33γ(u): A = -0.35 ± 0.05	Ap57
				1.17γ, 1.33γ(u): A = +0.295 ± 0.054; accurate indication that nonconservation of parity and noninvariance under charge conjugation are maximum	Ap58
		C_f	β ~150° 1.17γ,1.33γ(u)	1.17γ, 1.33γ(u): A = -0.344 ± 0.09	De57
		C_f	β ~150° 1.17γ,1.33γ(u)	1.17γ, 1.33γ(u): A = -0.34 ± 0.02	St58
Zn^{65}		C_t	n polarization 0°,180° 7.88γ	0.052 state: 3/2 (Cr54)	Tr57

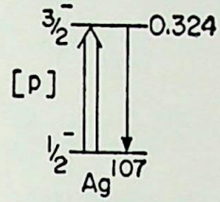
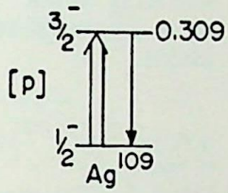
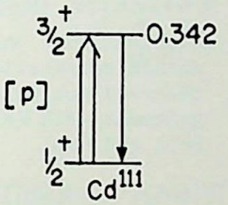
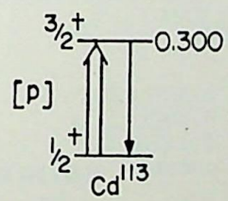
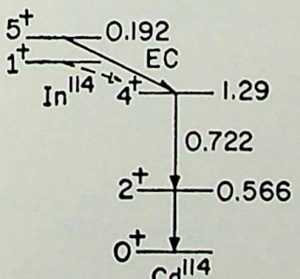
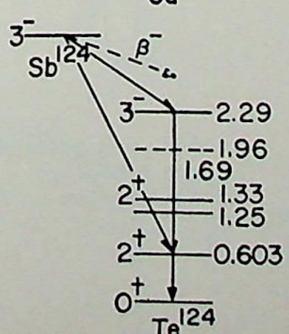
CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

TABLE XIII (*continued*).

Nucleus	Source or Reaction	Polarization Technique	Experimental Conditions	Summary of Results	Reference
Se^{76}		C	β 90° 0.55γ	0.55γ: E2 (Ri52, Wa50a)	Ha53a
Sr^{86}		C	β 90° 1.08γ	1.08γ:E2 (St51,Ri52)	Ha53a
Sr^{88}		C	1.85γ 90° 0.91γ	1.85γ: E2, 0.91γ: E1; cascade: 3(E1)2(E2)0 agrees with Pe48 and Me52a	Bi55
		C	1.85γ; 0.91γ 90° → 180° 0.91γ; 1.85γ	0.91γ:E1 1.85γ:E2 agrees with Bi55	Co56
Nb^{95}		C_f	β 153° 0.722γ, 0.754γ(u)	0.722γ, 0.754γ(u): A = -0.46 ± 0.09; S,T or V,A interference is present	Ap57
Mo^{95}		C	p 90° 0.203γ	0.203γ: $(E2/M1)^{\frac{1}{2}}$ = -0.58 ± 0.20	Mc58
Rh^{103}		C	p 90° 0.298γ	0.298γ: $(E2/M1)^{\frac{1}{2}}$ = -0.17	Mc58
Pd^{106}		C	γ 90° → 180° 0.513γ, 0.73γ(u)	0.513γ: E2, 0.73γ: E2: cascade: 0(E2)2(E2)0 tentatively (Br48)	Me50

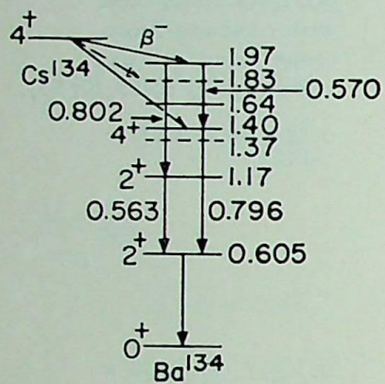
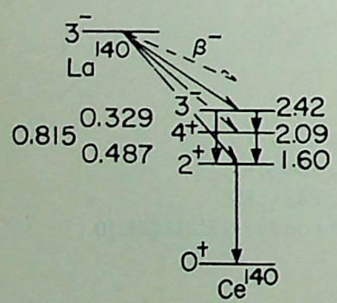
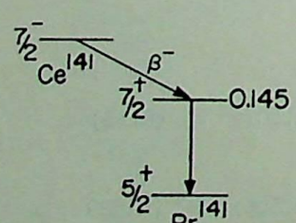
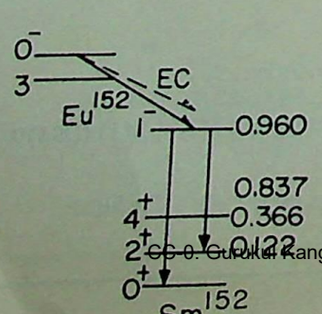
CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

TABLE XIII (*continued*).

Nucleus	Source or Reaction	Polarization Technique	Experimental Conditions	Summary of Results	Reference
Pd^{106}		C	γ $90^\circ \rightarrow 180^\circ$ 0.513γ, 0.73γ(u)	0.513γ: E2; cascade: suggest 0(E2)2(E2)0 mixed with cascade from close lying 3rd exc. state (Sp50, K153, Kr53).	Wi50
Ag^{107}		C	p 90° 0.324γ	0.324γ: $(E2/M1)^{\frac{1}{2}}$ = -0.21	Mc58
Ag^{109}		C	p 90° 0.309γ	0.309γ: $(E2/M1)^{\frac{1}{2}}$ = -0.19	Mc58
Cd^{111}		C	p 90° 0.342γ	0.342γ: $(E2/M1)^{\frac{1}{2}}$ = 0.385	Mc58
Cd^{113}		C	p 90° 0.300γ	0.300γ: $(E2/M1)^{\frac{1}{2}}$ = 0.29	Mc58
Cd^{114}		C	γ $90^\circ \rightarrow 270^\circ$ 0.566γ, 0.722γ(u)	0.566γ:E2 0.722γ:E2 cascade:4(E2)2(2E)0	Br56
Te^{124}		C	β 90° 0.605γ	0.605γ: E2 (St51)	St52
		C	β $90^\circ \rightarrow 180^\circ$ 0.605γ	0.605γ: E2 (K152)	K152
		C	γ $90^\circ \rightarrow 180^\circ$ 1.69γ,0.605γ(u)	0.605γ: E2, 1.69γ: E2/M1 = 0.09, cascade:3(E2, M1)2(E2)0	K152
		C	β 90° 0.605γ	0.605γ: E2 (Be50b)	Ha53a

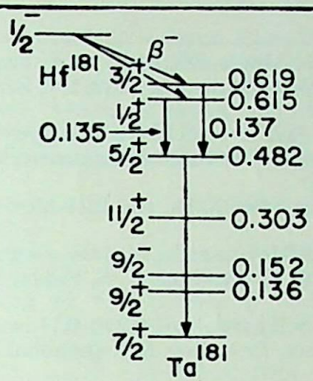
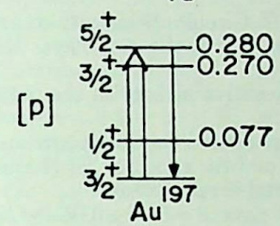
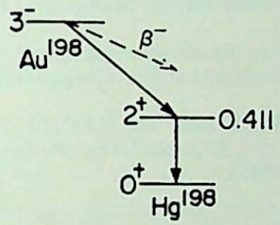
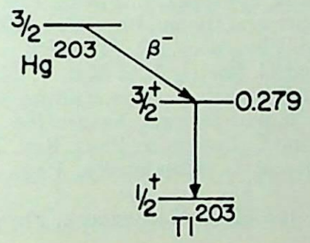
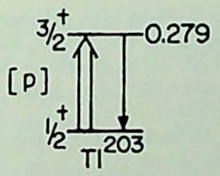
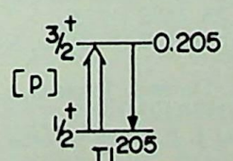
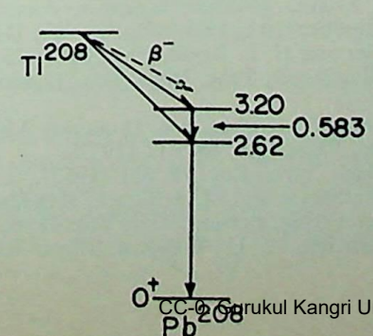
CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

TABLE XIII (*continued*).

Nucleus	Source or Reaction	Polarization Technique	Experimental Conditions	Summary of Results	Reference
Te^{124}		C_f	β 153° 0.605γ	0.605γ: A = 0.13 ± 0.06	Ap57
Ba^{134}		C	γ 90° → 180° 0.563γ,0.570γ, 0.605γ,0.796γ(u)	polarization consistent with a composite correlation coefficient = 0.13	Me50
		C	γ 90° → 180° 0.563γ,0.570γ, 0.605γ,0.796γ(u)	γ: unpolarized; see K152 which agrees with Me50 (St55)	Wi50
		C	γ 90° → 180° 0.563γ,0.570γ, 0.605γ,0.796γ(u)	0.605γ: E2, 0.796γ: E2; 1.402 - 0.605 cascade: 4(E2)2(E2)0, (Ro51)	K152
		C	γ polarimeter 180° 0.563γ,0.570γ, 0.605γ,0.796γ(u)	polarization - polarization correlation observed not in disagreement with Wa50, Pe51 and Ro51	Ro52
		C	β 90° 0.563γ,0.570γ, 0.605γ,0.796γ(u)	0.605γ, 0.796γ(u): unpolarized (Be50a, St50)	Ha53a
Ce^{140}		C	1.60γ 90° 0.329γ;0.487γ; 0.815γ	0.329γ: E1, 0.487γ: E2, 0.815γ: E1; cascade: 3(E1)4(E2)2(E2)0	Bi55
Pr^{141}		C	crystal electric field 90° 0.145γ	0.145γ: E2/M1 = 0.08 ± 0.02 (Ca55)	Ca55
Sm^{152}		C_t	ν 167° 0.960γ	0.960γ: neg.helicity, ν: neg.helicity; GT interaction is A,	Go58

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

TABLE XIII (*continued*).

Nucleus	Source or Reaction	Polarization Technique	Experimental Conditions	Summary of Results	Reference
Ta^{181}		C	0.133γ 90° 0.482γ	0.482γ: $(E2/M1)^{\frac{1}{2}} = 6.3$; 1.33 - 0.482 cascade: $1/2^+ - 5/2^+ - 7/2^+$ (Pa55, Me54a, He55)	St57
Au^{197}		C	p 90° 0.280γ	0.280γ: $(E2/M1)^{\frac{1}{2}} = -0.41$ ±0.04	Mc58
Hg^{198}		C_f	β 159° 0.411γ	0.411γ: A = +0.52 ± 0.09; max. interference between different T (or A) and S (or V) matrix elements	Bo57 Bo57a Bo58
		C_f	β ~150° 0.411γ	0.411γ: A = 0.34 ± 0.05	Be57a
Tl^{203}		C_f	β 159° 0.279γ	0.279γ: A = -0.06 ± 0.22	Bo57 Bo58
		C	p 90° 0.279γ	0.279γ: $(E2/M1)^{\frac{1}{2}} = 1.50$ ± 0.08	Mc58
Tl^{205}		C	p 90° 0.205γ	0.205γ: $(E2/M1)^{\frac{1}{2}} = 1.46$ ± 0.16	Mc58
Pb^{208}		C	γ 90° 0.583γ, 2.61γ(u)	0.583γ: E2, 2.61γ: E2; cascade: 4(E2)2(E2)0 (Pe50)	Kr53

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

TABLE XIII (addenda). Summary of experiments. An explanation of the various entries is given in Part III, E.

Nucleus	Method	Addenda. In the second column the letters have the same meaning as before. Result and reference
N^{14}	D	Measurement of linear polarization of γ ray in $C^{13}(p,\gamma)N^{14}$ at $E_p=1.76$ Mev indicated no parity change in the transition to the ground state [Strassenburg, Hubert, Krone, and Prosser, Bull. Am. Phys. Soc. Ser. II, 3, 372 (1958)].
Mg^{24}	C_f	Measurement of β-γ circular-polarization correlation in decay of Na^{24} gave $A=+0.05\pm0.04$ in agreement with other results [R. M. Steffen and P. Alexander, reported in *Proceedings of the Rehoveth Conference on Nuclear Structure* (North-Holland Publishing Company, Amsterdam, 1958)].
P^{31}	D	Linear polarization measured at three resonances in $Si^{30}(p,\gamma)P^{31}$. The 8.04-, 8.20-, and 8.24-Mev γ rays are $E1$, $M1$, and $E1$, respectively (Tu 57).
S^{34}	C	Measurement of γ-γ direction-polarization correlation in the 1.16–2.10-Mev cascade, with the γ-rays resolved, confirmed the assignment of $2^+(M1,E2)2^+(E2)0^+$ and gave $\delta=0.13$ [C. G. Shute and P. S. Fisher, Phil. Mag. 3, 726 (1958)]
Ti^{46}	C_t	Measurement of β-γ circular-polarization correlation in the decay of Sc^{46} gave $A=+0.29\pm0.11$ in agreement with other results. [Lundby, Patro, and Stroot, quoted by M. A. Grace, Proc. Roy. Soc. (London) A246, 460 (1958).]
Ni^{60}	C_t	The β-γ circular-polarization correlation was measured as a function of v/c. Circular polarizations greater than $\frac{1}{3}v/c$ were observed, especially for small values of v/c [Page, Pettersson, and Lindqvist, Phys. Rev. (to be published)].
Sm^{152}	C_t	The experiment of Go 58 was repeated to show that the neutrino has negative helicity in the decay of Eu^{152} [I. Marklund and L. A. Page, Nuclear Phys. 9, 88 (1958)].
W^{183}	C_t	Measurement of circular-polarization of radiation following capture of polarized thermal neutrons indicated complete right circular polarization in agreement with spin assignments and the expected $E1$ character of the radiation. Result used by Tr 57 to establish the reliability of experimental arrangement.
Hg^{195}	C_f	Measurement of β-γ circular-polarization correlation on decay of Au^{198} gave $A=0.45\pm0.06$ for an electron energy range of 520–620 kev and $A=0.38\pm0.10$ for a range of 620–850 kev (R. M. Steffen and P. Alexander, see Mg^{24} above).
Hg^{198}	C	The 411-kev γ ray from Au^{198} was resonant scattered by Hg^{198} and the linear polarization of the scattered radiation was measured. A multipolarity of $E2$ was confirmed for the radiation [V. Knapp and B. S. Sood, Proc. Roy. Soc. (London) A247, 369 (1958)].
Pb^{208}	C	Linear polarization measurements [G. T. Wood and P. S. Jastram, Phys. Rev. 100, 1237(A) (1955)] along with directional measurements confirmed the assignments, of 0^+, 3^-, 5^-, 4^-, and 5^- to the first 5 states in Pb^{208} [Elliott, Graham, Walker, and Wolfson, Phys. Rev. 93, 356 (1954)].

the direction-direction correlation) if it seems particularly relevant or related to the polarization measurement. No attempt is made at completeness in the listing of relevant references.

Column VI. Reference.—The reference corresponding to each experiment listed in the table is given in this column.

ACKNOWLEDGMENTS

We would like to thank Dr. Joachim Ehrman for many helpful discussions and Dr. Francis E. Throw for his generous aid and advice in writing the manuscript. Dr. Phillip Malmberg, Mr. Ronald Fast, and Mr. Samuel Chappell were of great assistance in the search of the literature. Mr. Chappell also helped with some of the calculations. Mr. Ralph Klingler provided many of the calculations on which the tables are based. The Graphic Arts Branch of NRL supplied most of the figures and reproductions. We are grateful also to Dorothy Bodden for her careful preparation of the manuscript. One of us (SSH) would like to thank the Guggenheim Memorial Foundation for a fellowship which aided in the work.

BIBLIOGRAPHY

Ag 57 A. Agodi, Nuovo cimento **5**, 21 (1957).
Al 54 D. E. Alburger and M. A. Grau, Proc. Phys. Soc. (London) **A67**, 280 (1954).
Al 55 L. Allen, Jr., Phys. Rev. **98**, 705 (1955).
Al 56 Alder, Bohr, Huus, Mottelson, and Winther, Revs. Modern Phys. **28**, 432 (1956).
Al 57 Alder, Stech, and Winther, Phys. Rev. **107**, 728 (1957).
Am 56 Ambler, Hudson, and Temmer, Phys. Rev. **101**, 196 (1956).
Am 57 Ambler, Hayward, Hoppes, and Hudson, Phys. Rev. **105**, 1413 (1957).
Ap 57 H. Appel and H. Schopper, Z. Physik **149**, 103 (1957).
Ap 58 Appel, Schopper, and Bloom, Phys. Rev. **109**, 2211 (1958).
Ar 50 W. R. Arnold, Phys. Rev. **80**, 34 (1950).
Ar 53 P. Argyres and G. Kittel, Acta Met. **1**, 241 (1953).
Au 27 P. Auger and F. Perrin, J. phys. radium **8**, 93 (1927).
Ba 50 Barnes, French, and Devons, Nature **166**, 145 (1950).
Be 50 H. A. Bethe and C. Longmire, Phys. Rev. **77**, 647 (1950).
Be 50a T. H. Berlin and L. Mandansky, Phys. Rev. **78**, 623 (1950).
Be 50b J. R. Beyster and M. L. Wiedenbeck, Phys. Rev. **79**, 728 (1950).
Be 56 H. A. Bethe and P. Morrison, *Elementary Nuclear Theory* (John Wiley and Sons, Inc., New York, 1956).
Be 57 D. B. Beard and M. E. Rose, Phys. Rev. **108**, 164 (1957).
Be 57a Berthier, Debrunner, Kündig, and Zwahlen, Helv. Phys. Acta **30**, 483 (1957).
Be 58 Benczer, Koller, Schwarzchild, Vise, and Wu, Phys. Rev. (to be published).
Bi 51 Biedenharn, Rose and Arfkin, Phys. Rev. **83**, 683 (1951).
Bi 52 Bishop, Daniels, Goldschmidt, Halban, Kurti, and Robinson, Phys. Rev. **88**, 1432 (1952).
Bi 53 L. C. Biedenharn and M. E. Rose, Revs. Modern Phys. **25**, 729 (1953).
Bi 54 Bishop, Daniels, Durand, Johnson, and Perez y Jorba, Phil. Mag. **45**, 1197 (1954).
Bi 55 G. R. Bishop and J. P. Perez y Jorba, Phys. Rev. **98**, 89 (1955).
Bi 55a Bishop, Grace, Johnson, Knipper, Lemmer, Perez y Jorba, and Scurlock, Phil. Mag. **46**, 951 (1955).
Bl 48 E. Bleuler and H. L. Bradt, Phys. Rev. **73**, 1398 (1948).
Bl 51 R. J. Blin-Stoyle, Proc. Phys. Soc. (London) **A64**, 700 (1951).
Bl 51a B. Bleaney, Proc. Phys. Soc. (London) **A64**, 315 (1951).
Bl 52 S. Bloom, Phys. Rev. **88**, 312 (1952).
Bl 53 C. F. Black, Phys. Rev. **90**, 381 (1953).
Bl 57 R. J. Blin-Stoyle and M. A. Grace, *Handbuch der Physik* (Springer Verlag, Berlin, 1957), Vol. 42, p. 555.
Bo 57 F. Boehm and A. H. Wapstra, Phys. Rev. **106**, 1364 (1957).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

Bo 57a F. Boehm and A. H. Wapstra, Phys. Rev. **107**, 1462 (1957).
Bo 57b F. Boehm and A. H. Wapstra, Phys. Rev. **107**, 1202 (1957).
Bo 58 F. Boehm and A. H. Wapstra, Phys. Rev. **109**, 456 (1958).
Bo 58a F. Boehm, Phys. Rev. **109**, 1018 (1958).
Br 48 E. L. Brady and M. Deutsch, Phys. Rev. **74**, 1541 (1948)
Br 54 Bretscher, Alderman, Elwyn, and Shull, Phys. Rev. **96**, 103 (1954).
Br 56 J. N. Brazos and R. M. Steffen, Phys. Rev. **102**, 753 (1956).
Br 57 Brini, Peli, Rimondi, and Veronesi, Nuovo cimento **6**, 98 (1957).
Bu 24 F. W. Bubb, Phys. Rev. **23**, 137 (1924).
Ca 55 Cacho, Grace, Johnson, Knipper, Scurlock, and Taylor, Phil. Mag. **46**, 1287 (1955).
Ca 57 Cavanagh, Turner, Coleman, Gard, and Ridley, Phil. Mag. **21**, 1105 (1957).
Cl 52 F. P. Clay and F. L. Hereford, Phys. Rev. **85**, 675 (1952).
Co 53 F. Coester and J. M. Jauch, Helv. Phys. Acta **26**, 3 (1953).
Co 56 C. F. Coleman, Phil. Mag. **1**, 166 (1956).
Co 56a C. F. Coleman, Phys. Rev. **103**, 647 (1956).
Cr 54 B. Crasemann, Phys. Rev. **93**, 1034 (1954).
Da 52 Daniels, Grace, Halban, Kurti, and Robinson, Phil. Mag. **43**, 1297 (1952).
Da 52a C. M. Davisson and R. D. Evans, Revs. Modern Phys. **24**, 79 (1952).
Da 53 Daniel, Koester, and Mayer-Kuchuk, Z. Naturforsch. **8a**, 447 (1953).
De 50 M. Deutsch and K. Siegbahn, Phys. Rev. **77**, 680 (1950).
De 57 P. Debrunner and W. Kündig, Helv. Phys. Acta **30**, 261 (1957).
De 57a S. Devons and L. J. B. Goldfarb, *Handbuch der Physik* (Springer Verlag, Berlin, 1957), Vol. 42, p. 362.
De 58 J. J. deSwart and R. E. Marshak, Phys. Rev. (to be published).
Di 58 Diddens, Huiskamp, Severiens, Miedema, and Steenland, Nuclear Phys. **5**, 58 (1958).
Es 56 Estulin, Popov, and Chukreev, Zhur. Eksptl. i Teort. Fiz. **30**, 1052 (1956).
Ev 55 R. D. Evans, *The Atomic Nucleus* (McGraw-Hill Book Company, Inc., New York, 1955).
Fa 48 D. L. Falkoff, Phys. Rev. **73**, 518 (1948).
Fa 49 U. Fano, J. Opt. Soc. Am. **39**, 859 (1949).
Fa 53 L. W. Fagg and S. S. Hanna, Phys. Rev. **92**, 372 (1953).
Fa 54 U. Fano, Phys. Rev. **93**, 121 (1954).
Fe 55 M. Ferentz and N. Rosenzweig, Argonne National Laboratory Report ANL-5324 (unpublished).
Fi 31 J. Fischer, Ann. Physik **8**, 821 (1931).
Fo 49 Fowler, Lauritsen, and Toolestrup, Phys. Rev. **76**, 1767 (1949).
Fr 38 W. Franz, Ann. Physik **33**, 698 (1938).
Fr 52 A. P. French and J. O. Newton, Phys. Rev. **85**, 1041 (1952).
Fu 50 E. G. Fuller, Phys. Rev. **79**, 303 (1950).
Ga 57 Yu. V. Capanov and V. S. Popov, Zhur. Eksptl. i Teort. Fiz. **33**, 256 (1957).
Go 34 C. G. Gorter, Physik. Z. **35**, 923 (1934).
Go 46 G. Goertzel, Phys. Rev. **70**, 897 (1946).
Go 48 C. J. Gorter, Physica **14**, 504 (1948).
Go 51 G. Goldhaber, Phys. Rev. **81**, 930 (1951).
Go 58 Goldhaber, Grodzins, and Sunyar, Phys. Rev. **109**, 1015 (1958).
Gr 54 Grace, Johnson, Kurti, Lemmer, and Robinson, Phil. ag. **45**, 1192 (1954).
Gr 55 Grant, Rutherglen, Flack, and Hutchinson, Proc. Phys. Soc. (London) **A68**, 369 (1955).
Gr 55a T. P. Gray and G. R. Satchler, Proc. Phys. Soc. (London) **A68**, 349 (1955).
Gr 56 D. F. Griffing and J. C. Wheatley, Phys. Rev. **104**, 389 (1956).
Gu 53 S. B. Gunst and L. A. Page, Phys. Rev. **92**, 970 (1953).
Ha 40 D. R. Hamilton, Phys. Rev. **58**, 122 (1940).
Ha 48 D. R. Hamilton, Phys. Rev. **74**, 782 (1948).
Ha 48a R. C. Hanna, Nature **162**, 332 (1948).
Ha 51 O. Halpern, Phys. Rev. **82**, 753 (1951).
Ha 51a O. Halpern, Nature **168**, 782 (1951).
Ha 53 J. Halpern and E. V. Weinstock, Phys. Rev. **91**, 934 (1953).
Ha 53a Hamilton, Lemonik, and Pipkin, Phys. Rev. **92**, 1191 (1953).
He 44 W. Heitler, *The Quantum Theory of Radiation* (Oxford University Press, London, 1944).
He 51 F. L. Hereford, Phys. Rev. **81**, 482 (1951).
He 51a F. L. Hereford, Phys. Rev. **81**, 627 (1951).
He 53 F. L. Hereford and J. P. Keuper, Phys. Rev. **90**, 1043 (1953).
He 55 Heer, Ruetschi, Gimmi, and Kundig, Helv. Phys. Acta **28**, 336A (1955).
Ho 53 J. R. Holt and T. U. Marsham, Proc. Phys. Soc. (London) **A66**, 467, 565 (1953).
Hu 55 Hurley, Rudman, and Jastram, Phys. Rev. **98**, 1187(A) (1955).
Hu 56 I. S. Hughes and D. Sinclair, Proc. Phys. Soc. (London) **A69**, 125 (1956).
Hu 56a Huiskamp, Steenland, Miedema, Tolhoek, and Gorter, Physica **22**, 587 (1956).
Hu 57 Huiskamp, Diddens, Severiens, Miedema, and Steenland, Physica **23**, 605 (1957).
Hu 57a L. Hulthén and M. Sugaware, *Handbuch der Physik* (Springer Verlag, Berlin, 1957), Vol. 39, p. 129.
Ka 58 Katz, Kowalski, Brown, and Baerg, Bull. Am. Phys. Soc. Ser. II, **3**, 174 (1958).
Ke 54 G. L. Keister, Phys. Rev. **96**, 855(A) (1954).
Kh 55 G. P. Khutsishvili, J. Exptl. Teoret. Phys. (U.S.S.R.) **28**, 496 (1955).
Ki 31 P. Kirkpatrick, Phys. Rev. **38**, 1938 (1931).
Kl 29 O. Klein and Y. Nichina, Z. Physik **52**, 853 (1929).
Kl 52 Kloeper, Lennox, and Wiedenbeck, Phys. Rev. **88**, 695 (1952).
Kl 53 E. D. Klema and F. K. McGowan, Phys. Rev. **92**, 1469 (1953).
Ko 56 O. Kofoed-Hansen and A. Winther, Kgl. Danske Videnskab. Selskab, Mat.-fys. Medd **30**, No. 20 (1956).
Kr 53 J. J. Kraushaar and M. Goldhaber, Phys. Rev. **89**, 1081 (1953).
Ku 35 N. Kurti and F. E. Simon, Proc. Roy. Soc. (London) **A149**, 152 (1935).
Le 55 Lemmer, Segaert, and Grace, Proc. Phys. Soc. (London) **A68**, 701 (1955).
Le 56 T. D. Lee and C. N. Yang, Phys. Rev. **104**, 254 (1956).
Le 57 T. D. Lee and C. N. Yang, Phys. Rev. **105**, 1671 (1957).
Li 49 D. S. Ling and D. L. Falkoff, Phys. Rev. **76**, 1639 (1949).
Li 54 F. W. Lipps and H. A. Tolhoek, Physica **20**, 85 (1954).
Li 54a F. W. Lipps and H. A. Tolhoek, Physica **20**, 395 (1954).
Li 58 A. E. Litherland and H. E. Gove, Bull. Am. Phys. Soc. Ser. II, **3**, 200 (1958).
Ll 52 S. P. Lloyd, Phys. Rev. **88**, 906 (1952).
Lu 57 Lundby, Patro and Stroot, Nuovo cimento **6**, 745 (1957).
Ma 49 J. F. Marshall and E. Guth, Phys. Rev. **76**, 1879 (1949).
Ma 50 J. F. Marshall and E. Guth, Phys. Rev. **78**, 738 (1950).
Ma 51 J. E. MacDonald and D. L. Falkoff, Phys. Rev. **83**, 875 (1951).
Ma 58 F. Mandl, Proc. Phys. Soc. (London) **A71**, 177 (1958).
Mc 54 W. H. McMaster and F. L. Hereford, Phys. Rev. **95**, 723 (1954).
Mc 54a F. K. McGowan, Phys. Rev. **93**, 471 (1954).
Mc 58 F. K. McGowan and P. H. Stelson, Phys. Rev. **109**, 901 (1958).
Mc 58a C. L. McGinnis (private communication).
Me 50 F. Metzger and M. Deutsch, Phys. Rev. **78**, 551 (1950).
Me 52 F. R. Metzger, Phys. Rev. **86**, 435 (1952).
Me 52a F. R. Metzger and H. C. Amacher, Phys. Rev. **88**, 147 (1952).
Mo 57 M. Morita and R. S. Morita, Phys. Rev. **107**, 1316 (1957).
My 38 R. D. Myers, Phys. Rev. **54**, 361 (1938).
Ni 29 Y. Nishina, Z. Physik **52**, 869 (1929).
Pa 55 H. Paul and R. M. Steffen, Phys. Rev. **98**, 231 (1955).
Pe 48 W. C. Peacock and J. W. Jones, U. S. Atomic Energy Commission Declassified Document AECD 1812, March 1948 (unpublished).
Pe 50 H. E. Petch and M. W. Johns, Phys. Rev. **80**, 478 (1950).
Pe 51 C. L. Peacock and J. L. Braud, Phys. Rev. **83**, 484 (1951).
Po 49 R. V. Pound, Phys. Rev. **76**, 1410 (1949).
Po 54 O. J. Poppema, thesis, Groningen (1954).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

Po 57 Postma, Huiskamp, Miedema, Steenland, Tolhoek, and Gorter, Physica **23**, 259 (1957).
Pr 47 M. H. L. Pryce and J. C. Ward, Nature **160**, 435 (1947).
Pr 54 W. W. Pratt, Phys. Rev. **95**, 1517 (1954).
Ri 52 S. L. Ridgway and F. M. Pipkin, Phys. Rev. **87**, 202 (1952).
Ro 49 M. E. Rose, Phys. Rev. **75**, 213 (1949).
Ro 51 B. L. Robinson and L. Madansky, Phys. Rev. **84**, 604 (1951).
Ro 52 B. L. Robinson and L. Madansky, Phys. Rev. **88**, 1065 (1952).
Ru 54 Rutherglen, Grant, Flack, and Deucbars, Proc. Phys. Soc. (London) **A67**, 101 (1954).
Ru 56 Rutherglen, Deucbars, and Wallace (private communication in Hu 56).
Sa 31 P. Sauter, Ann. Physik **11**, 454 (1931).
Sa 51 E. E. Salpeter, Phys. Rev. **82**, 60 (1951).
Sa 52 J. E. Sanders, Phil. Mag. **43**, 630 (1952).
Sa 53 G. R. Satchler, Proc. Phys. Soc. (London) **A66**, 1081 (1953).
Sa 54 G. R. Satchler, Phys. Rev. **94**, 1304 (1954).
Sa 55 F. Sauter and H. O. Wuster, Z. Physik **141**, 83 (1955).
Sa 55a M. Sakai, J. Phys. Soc. Japan **10**, 729 (1955).
Sa 55b G. R. Satchler, Proc. Phys. Soc. (London) **A68**, 1041 (1955).
Sc 30 G. Schur, Ann. Physik **4**, 433 (1930).
Sc 50 L. I. Schiff, Phys. Rev. **78**, 733 (1950).
Sc 57 H. Schopper, Phil. Mag. **2**, 710 (1957).
Se 52 J. Seed and A. P. French, Phys. Rev. **88**, 1007 (1952).
Se 53 E. Segre, *Experimental Nuclear Physics* (John Wiley & Sons, Inc., New York, 1956), Vol. I.
Si 51 Simon, Rose, and Jauch, Phys. Rev. **84**, 1155 (1951).
Si 53 A. Simon and T. A. Welton, Phys. Rev. **90**, 1036 (1953).
Si 53a A. Simon, Phys. Rev. **92**, 1050 (1953).
Sn 48 Snyder, Pasternak and Hornbostel, Phys. Rev. **73**, 440 (1948).
So 30 A. Sommerfeld and G. Schur, Ann. Physik **4**, 409 (1930).
Sp 50 J. A. Spiers, Phys. Rev. **78**, 75 (1950).
St 50 R. Stump and S. Frankel, Phys. Rev. **79**, 243 (1950).
St 51 D. T. Stevenson and M. Deutsch, Phys. Rev. **84**, 1071 (1951).
St 51a R. M. Steffen, Phys. Rev. **83**, 166 (1951).
St 52 R. Stump, Phys. Rev. **86**, 249 (1952).
St. 52a N. R. Steenberg, Proc. Phys. Soc. (London) **A65**, 791 (1952).
St 53 Steenberg, Proc. Phys. Soc. (London) **A66**, 391 (1953).
St 53a N. R. Steenberg, Proc. Phys. Soc. (London) **A66**, 399 (1953).
St 53b N. R. Steenberg, Can. J. Phys. **31**, 204 (1953).
St 55 Stewart, Scharenberg, and Wiedenbeck, Phys. Rev. **99**, 691 (1955).
St 56 R. M. Steffen and J. N. Brazos, Phys. Rev. **99**, 1646(A) (1956).
St 57 P. H. Stelson and F. K. McGowan, Phys. Rev. **105**, 1346 (1957).
St 57a M. J. Steenland and H. A. Tolhoek, *Progress in Low Temperature Physics* (North Holland Publishing Company, Amsterdam, 1957), Vol. II, p. 292.
St 58 R. M. Steffen, Bull. Am. Phys. Soc. Ser. II, **3**, 205 (1958).
St 58a R. M. Steffen (private communication).
To 52 H. A. Tolhoek and J. A. M. Cox, Physica **18**, 357 (1952).
To 52a H. A. Tolhoek, Physica **18**, 1257 (1952).
To 53 H. A. Tolhoek and J. A. M. Cox, Physica **19**, 101 (1953).
To 56 H. A. Tolhoek, Revs. Modern Phys. **28**, 277 (1956).
Tr 57 G. Trumpy, Nuclear Phys. **2**, 664 (1957).
Tr 57a G. Trumpy (to be published).
Tu 57 Tutakin, Tzytco, Gonchar, Walter, and Lwov, USSR Conference on Low-Energy Nuclear Reactions, Moscow (1957).
Va 56 J. H. Van Vleck, Bull. Am. Phys. Soc. Ser. II, **1**, 11 (1956).
Vl 49 Vlasov, Dzhelepov, Doklady Akad. Nauk. **69**, No. 6, 777 (1949).
Wa 50 Waggoner, Moon, and Roberts, Phys. Rev. **80**, 420 (1950).
Wa 50a Walter, Huber, and Zunti, Helv. Phys. Acta **23**, 697 (1950).
Wh 46 J. A. Wheeler, Ann. N. Y. Acad. Sci. **48**, 219 (1946).
Wh 55 Wheatley, Huiskamp, Diddens, Steenland, and Tolhoek, Physica **21**, 841 (1955).
Wh 58 A. Whetstone and J. Halpern, Phys. Rev. **109**, 2072 (1958).
Wi 23 C. T. R. Wilson, Proc. Roy. Soc. (London) **A104**, 11 (1923).
Wi 50 A. H. Williams and M. L. Wiedenbeck, Phys. Rev. **78**, 822 (1950).
Wi 51 G. C. Wick, Phys. Rev. **81**, 467 (1951).
Wi 51a M. Wiener, National Bureau of Standards Circular 515 (1951).
Wi 52 D. H. Wilkinson, Phil. Mag. **43**, 659 (1952).
Wi 56 E. J. Winhold and I. Halpern, Phys. Rev. **103**, 990 (1956).
Wi 58 D. H. Wilkinson, Phys. Rev. **109**, 1610 (1958).
Wo 49 L. Wolfenstein, Phys. Rev. **75**, 1664 (1949).
Wo 55 G. T. Wood and P. S. Jastram, Phys. Rev. **98**, 1187(A) (1955).
Wu 50 C. S. Wu and I. Shaknov, Phys. Rev. **77**, 136 (1950).
Wu 57 Wu, Ambler, Hayward, Hoppes, and Hudson, Phys. Rev. **105**, 1413 (1957).
Zi 50 I. Zinnes, Phys. Rev. **80**, 386 (1950).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

Electron and Positron Polarization

LORNE A. PAGE

University of Pittsburgh, Pittsburgh, Pennsylvania

Table of Contents

I. Introduction

II. General Considerations
- A. Polarization of Free Electrons
- B. Polarization Efficiencies
- C. Right *versus* Left Helicity

III. Experimental Methods
- A. Classification of Methods
- B. Mott Scattering (and Spin Precessors)
- C. Energy Term $\sigma \cdot \mathbf{H}$
- D. Comparison Methods
  - (1) Møller and Bhabha Scattering
  - (2) Annihilation of Slow Positrons in a Ferromagnet
  - (3) Annihilation in Flight in Polarized Electron Target
  - (4) Polarized Bremsstrahlung
  - (5) Circular Polarization of Annihilation Photons

IV. Concluding Remarks
- A. On Calculations
- B. On Experiment
- C. On Calibration

V. Acknowledgments

Appendix A. Beta Decay and Polarization

Appendix B. Short Compendium of Helicity Signs

Appendix C. Compton Analysis of Photon Helicity

Bibliography

I. INTRODUCTION

WHEN Goudsmit and Uhlenbeck proposed the spinning electron the possibility arose of having spin-polarization effects for a free electron. On the basis of Dirac's theory of the electron Mott proposed in 1929 a double scattering experiment designed to demonstrate such (transverse) polarization: the first scattering in the Coulomb field of a high-Z nucleus would serve partially to polarize an incident beam of unpolarized electrons because of the spin-orbit interaction, while the second such scattering would serve as the polarization analyzer. For a quarter century thereafter this quite difficult experiment was pursued in a number of laboratories. Approximately midway through this period it became evident that the experiment could, in fact, demonstrate the sought-for polarization, and at present there remains no doubt whatever that it can be made to work.* It is perhaps fortunate that the double scattering experiment was not crucial for Dirac theory, and that the other verifications came as early as they did.

An unfortunate situation prevailed during the 1930's and 1940's in that the whole matter of electron, positron, and gamma† polarization in general was viewed rather apathetically by many physicists. Such attention as there was, was focused almost entirely on the double scattering experiment itself. Thus the polarization effects already existing in the Klein-Nishina matrix element for Compton scattering, in the Møller matrix element for electron-electron or positron-electron scattering, in the Bethe-Heitler matrix element for bremsstrahlung, to name a few of the common electron-photon processes, continued to lie dormant‡ until quite recently.

Prior to about 1956 there was no supposition that "prepolarized" beams of electrons, positrons, or photons would be available from beta decay of nuclei. Therefore any experiments then proposed, just as with the Mott proposal, would have had to be of "second order," involving a polarizing unit to be followed by an analyzing unit. Such experiments seem to have had no strong appeal to experimenters in general.

When it turned out that parity violation in beta decay was in a sense complete, as demonstrated by the beautiful cobalt-60 experiment of Wu, Ambler, Hayward, Hoppes, and Hudson, it became clear that prepolarized beams of particles (and at times prepolarized photons) must exist. This initiated a greatly accelerated development of practical techniques for handling such polarizations. This paper discusses the various methods for analyzing electron and positron polarization, most of which were introduced and proved during the short period of about half a year following the announcement of parity violation. Photon (circular) polarization§ must be brought into the discussion

* Tolhoek [reference (T)] covers the double scattering experiment thoroughly.

† Plane polarization was discussed early, but there is no *direct* connection between such photon polarization and electron polarization.

‡ That is, spin polarization was rigorously summed or averaged at each opportunity in proceeding from matrix element to transition probability. A rare exception is the paper of Franz (F38). The spin direction having been generally taken out, circular polarization information was automatically suppressed. The justification for such averaging was considered sound: "The spin direction is not measured." We have not far to look for the reason it was not measured, and the circle is thus closed.

§ Such polarization is dealt with only lightly here. In contrast to the situation regarding particle polarization, there appears to have been introduced no new approach to this problem since the time of the Wu experiment. Reference (T) discusses the Compton effect as applied to circular polarization measurement. See also a recent review (S58) by Schopper.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

because it connects with the problem of measuring free particle polarization.

The large and decisive effect found in the Wu experiment was primarily responsible for the apparent ease with which the new polarization methods succeeded. Knowledge that certain polarizations are quite probably present, and in abundance, has made for quick development of suitable experimental techniques. Also this knowledge has made for ready acceptance of most experimental conclusions relating to beta-decay polarizations, however hastily drawn a few may have been.

In the beginning of this new polarization era, one might have expected that only "first-order" experiments would appear. However, with the luxury of generally large inherent polarizations to begin with, a considerable number of second-order methods have already been successfully proved. Here the polarization one wishes to know is transferred for strategic reasons from initial particle to photon, or from particle to particle, in the first stage, and then the polarization analysis itself takes place in the second stage. Thus one requires two unlikely events in series, which makes the experiment second order.

The methods and the physics on which they are based are emphasized here, so the historical order is not followed in every case. For a given topic only a sufficient number of works is reviewed to illustrate the physics involved.‖ Closely related ideas (although perhaps not proved experimentally) are sometimes mentioned to round out the discussion. No pretense is made of reviewing all "parity results" obtained by a given method. No grand averages are made of the numbers obtained with a given radioactive source, and no pronouncements are attempted on such matters as a universal beta interaction or *PC* invariance.

The discussion cannot be divided rigidly into experimental method *versus* "theory of the method." Measurement of polarization by counter techniques (the only technique described here) is a field where it seems particularly safe to predict that a hypothetical ensemble of technicians continually trying different arrangements of scattering foils, magnets, absorbers, and so forth—with no belief in any theory and therefore no real plan—would take a long time indeed to happen upon a useful method. Each technique had to be planned in advance, hopefully at times because polarization measurement is not easy.

II. GENERAL CONSIDERATIONS

A. Polarization of Free Electrons

For an electron under no forces, solutions of the Dirac equation exist which are simultaneous eigenstates of momentum $\mathbf{p}$, of energy [eigenvalues $\pm\gamma$, where $\gamma=(1+p^2)^{\frac{1}{2}}$], and, if we wish, also of helicity. That is, letting $\mathbf{p}$ define the z direction, we may have a solution which is simultaneously as eigenstate of $H\equiv\boldsymbol{\alpha}\cdot\mathbf{p}+\beta$, of $\mathbf{p}$, and of σ_z (with eigenvalues ±1). This is possible because the operator σ_z commutes with all of the Hamiltonian *that matters* for this special case, namely, $p\alpha_z+\beta$.

But (still letting eigenvalue $\mathbf{p}$ define the z direction) we are not able to construct an eigenstate of σ_x, for σ_x anticommutes with α_z. However if the coefficient of α_z, namely p, is small, then the effect of this lack of commutation becomes correspondingly small. Thus at very low energy any component of $\boldsymbol{\sigma}$ is nearly completely measurable. For finite laboratory energy the closest we can come to having a state of precise σ_x is to take the "best" linear combination of the two eigenstates of σ_z. Mixing equal amplitudes of the latter two states with a plus (minus) sign gives $\langle\sigma_x\rangle=\pm1/\gamma$, which tends to ±1 as p tends to zero. Another real operator, $\beta\sigma_x$, does commute with α_z, and with β. Therefore, simply because the maximum expectation value of σ_x becomes so nearly zero at high energy, it does not follow that transverse polarization is somehow "unmeasurable" at high energies. This means only that techniques are required that aim at measuring $\beta\sigma_\perp$ rather than measuring $\sigma_\perp$ itself. Otherwise one might follow the course of first precessing the spin and then measuring σ_z.

In speaking of longitudinal polarization of an electron or positron we use $P_p=\langle\sigma_z\rangle$, where z is parallel to the velocity. P_p is thus a number which may range from -1 to $+1$. As is well known, a Lorentz transformation may be made along the z axis, to the rest system of the particle in particular, and P_p retains the same value. In the rest system we could for the moment suppose full polarization is some direction, say at angle θ_0 (polar angle) and ϕ_0 (azimuthal angle), referred to z and x directions. The spinor part of the state can then be given by $|+1\rangle\cos(\theta_0/2)+|-1\rangle e^{i\phi_0}\sin(\theta_0/2)$ in the rest system of the particle, or in the laboratory. For this situation the value of P_p is just $\cos\theta_0$.

Transverse polarization is denoted by P_s, and naturally a certain azimuth has to be implied. We shall mean that $P_s(\phi)=\langle\sigma_x\cos\phi+\sigma_y\sin\phi\rangle$ computed *in the rest system* of the particle. For the oblique situation given in the foregoing, $P_s(\phi)$ amounts to $\sin\theta_0\cos(\phi-\phi_0)$. Often in speaking of how much transverse polarization exists in a beam, one tacitly maximizes $|P_s|$ against ϕ before quoting the number, since the geometry of the experiment usually involves a natural choice for the reference azimuth ϕ_0 (or $\phi_0+\pi$).

Suppose we have a partially polarized beam (longitudinally polarized, for example), with $0<P<1$, say. We may consider that the fraction $1-|P|$ of the particles are unpolarized while the fraction $|P|$ are completely polarized, with the sign of P. Alternatively the same beam may be regarded as a mixture of + and

‖ The bibliography makes no attempt to list many of the excellent experiments which have corroborated, extended or improved upon those which are listed. The total literature on the subject (experimental and computational) has become quite extensive. Inasmuch as experimental technique in general seems now to be in an improvement stage rather than in a stage of new development, it seems too early to attempt complete listing.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

— polarized components in the ratio of $(1+P)/(1-P)$, which holds for either sign of P. Certain mathematical approaches favor the first view. On the other hand, most calculations proceed just as easily using the second view. Since no experimental detector can be devised to differentiate between the two situations, there is no point in debating which of these two views is the correct one.

To specify the total polarization in a beam it suffices to know three components of polarization in mutually orthogonal directions.¶ Polarization is averaged as follows: If beam number one of intensity m_1 has an amount of polarization $\mathbf{P}(1)$ and a second beam *incoherent* with the first, of intensity m_2, has $\mathbf{P}(2)$, then as a whole the polarization is

$$\mathbf{P}=[m_1\mathbf{P}(1)+m_2\mathbf{P}(2)]/(m_1+m_2).$$

B. Polarization Efficiencies

One is interested in the relative response of a polarization detector for two orthogonal spin orientations (that is, mutually orthogonal states). For example, for detecting a nonzero P_p one requires an apparatus that responds differently to a particle with $P_p=+1$ than to a particle with $P_p=-1$. All the detectors reviewed here are reversible.** Suppose for + setting of the detector that the probability, given a + particle, that it count is C_{++}, and the corresponding probability for a − particle to count is C_{+-}. Let the − setting of the detector have respective probabilities C_{-+} and C_{--}. The order of the subscripts is important. Then, if $C_{--}=C_{++}$ and also $C_{-+}=C_{+-}$, the detector is called *reversible*. We write $C_{++}/C_{+-}=r=C_{--}/C_{-+}$. (This definition of r is more useful than using C_{++}/C_{-+} or C_{--}/C_{+-} as appears in a moment.) The counting-rate ratio, + setting divided by − setting, will be $R=(1+\epsilon P)/(1-\epsilon P)$. Here the efficiency ϵ of the detector is $\epsilon=(r-1)/(r+1)$. This efficiency applies of course to one component of polarization, that associated with the ± setting of the detector.††

In practical circumstances, due to fringing fields, for example, a counting efficiency (not a polarization efficiency) may change from + setting to − setting, violating the condition for reversibility. But since the physics upon which the polarization sensitivity is based is independent of such systematic effects, it may still be that the ratio r as stated above is unambiguous. We then keep the same value of ϵ independent of the systematic effect. In such a case the counting-rate ratio has to be written more generally as $R=(C_{++}/C_{--})(1+\epsilon P)/(1-\epsilon P)$. If it should be possible exactly to reverse P, or to render $P=0$, say (and neither of these things may be easy to do), then the fact that $C_{++}/C_{--}=C_{+-}/C_{-+}$ is not unity need not hurt the final result, since this ratio can be determined empirically.

To keep the discussion simple, reversible detectors are assumed from now on. A good reason for generally thinking in terms of an efficiency ϵ instead of the corresponding ratio r is that the law of propagation for a number of "in series" effects is most clearly expressed in terms of ϵ's and not r's. For example (if we may adhere to the nonpositivistic approach), consider a particle or a beam which is polarized by an amount P_0. Suppose the polarization were transferred in the first stage of operations to some intermediate particle or beam with transfer efficiency ϵ_1, the intermediate polarization, perhaps of a different kind, being now $\epsilon_1 P_0$. Let this beam be now analyzed with efficiency ϵ_2. The over-all efficiency $\epsilon_1\epsilon_2$ is the quantity that defines the change in counting rate on reversing either P_0, ϵ_1, or ϵ_2. The rate ratio on such reversal is $R=(1+\epsilon_1\epsilon_2 P_0)/(1-\epsilon_1\epsilon_2 P_0)$. Depolarization could be taken into account by using a series of "efficiencies for preservation of polarization," numbers ϵ_d's less than +1 which would then multiply the $\epsilon_1\epsilon_2$ in this example. All of the ϵ's mentioned, including that of the detector, can as well be regarded as transfer efficiencies, since the over-all product-ϵ represents numerically the efficiency for transferring the original polarization information into a counting-rate asymmetry.

Here we have chosen to carry through a pair of *probabilities* rather than a pair of *amplitudes*. It has been tacitly assumed that physical judgement has been used throughout in choosing at each juncture a pair of orthogonal states for which any coherence between their amplitudes is of no consequence. The point is that the relative phase of their amplitudes is thrown away each time we force the calculation to tell us probability throughout the chain of events. Any blunder in such choice of states leads to grief. Often, there is considerable latitude here and we are free to choose a reasonably simple pair of states for algebraic convenience. If there is any doubt as to which pair of states to choose, which is tantamount to having doubt as to which kind of polarization it will suffice to talk about, then the amplitudes themselves could be carried through.

To return to the counting-rate asymmetry, experimenters usually prefer to discuss results in terms of the fractional change in rate between the + and − settings of the detector. This fractional change is often denoted by δ. One has

$$\delta=2\frac{R-1}{R+1}=2\frac{R_+-R_-}{R_++R_-}=2\epsilon P. \qquad (1)$$

¶ We thus have a polarization space of three dimensions for spin-$\frac{1}{2}$ particles. For photons (which are *not* spin-1 particles) a similar space can be used: The momentum direction $\mathbf{k}$ is special and pertains to circular polarization; normal to this, plane polarization may be represented, providing that azimuthal angle in this polarization space is twice the corresponding azimuthal angle in the laboratory.

** Actually, experimental situations exist and may play a role in the future where in practice the detector is not actually reversed. For example, it may have one state where it responds preferentially to $P>0$, and the other state where it is insensitive to P.

†† We could represent any polarization detector as a vector $\boldsymbol{\varepsilon}$ in a three-dimensional efficiency space. The efficiency for detecting arbitrary polarization $\mathbf{P}$ in a beam is then $\boldsymbol{\varepsilon}\cdot\mathbf{P}/|P|$. The same applies for photons.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

Concerning the factor 2 that appears in expression (1): given a pair of numbers, 100 and 101, one physicist may recognize and comment that they differ (to first order) by one percent. Another may prefer to state that these two numbers exhibit an asymmetry about their arithmetic mean of 1/201, thus a nominal one-half percent effect. We prefer the first method in discussing experimental results, even though δ will exceed 100% if experiments improve to the extent that ϵP becomes larger than 0.5.

It may be felt that density matrices and Stokes parameters are not accorded their fair place in the discussion. A number of review articles have treated these matters [T, (F57)]. The basic ideas of the experiments to be described can be put in quite simple language. This reviewer can readily believe (but would find it difficult to prove) that the majority of these experiments were conceived, made to work, and interpreted correctly, without conscious recourse to the density matrix.

Before ending the general discussion of polarization measurement we might inquire a little into measurability of more than one component of polarization. Consider some state α which means definite spin direction for the particle (in its rest system). Only one orthogonal state exists, for a spin-$\frac{1}{2}$ particle; call it β. Let us have α and β normalized in the same way, with the phase between them fixed. For any state of affairs, the polarization of the kind α is computed from the probability of finding that the particle is in state α compared to the probability it is in state β. Thus

$$P_\alpha=\frac{\text{prob}(\alpha)-\text{prob}(\beta)}{\text{prob}(\alpha)+\text{prob}(\beta)}, \qquad (2)$$

and P_β is $-P_\alpha$.

Instead of states α and β we might use a state called A, which is $\alpha+e^{i\phi}\beta$ and its orthogonal state B which is $\alpha+e^{i(\phi+\pi)}\beta$. Then the polarization of kind A, denoted P_A, is given in terms of prob(A) and prob(B) by an expression such as (2). So long as ϕ is a real number, polarization P_A is complementary to polarization P_α, since if $P_\alpha{}^2=1$, then $P_A{}^2=0$, and conversely. We may generate a third polarization $P_{A'}$ which is complementary to P_A and to $P_\alpha(A'=\alpha+ie^{i\phi}\beta, B'=\alpha+ie^{i(\phi+\pi)}\beta)$. For a particle polarized by an amount $\mathbf{P}$, we have

$$P_\alpha{}^2+P_A{}^2+P_{A'}{}^2=P^2\leqq 1. \qquad (3)$$

Of course, expression (3) is a statement of probabilities, as with expression (2). If one measures, by utterly precise means, P_A for a particle, one finds either $+1$ or -1, and expression (3) is then amended to read $P_A{}^2=1$, $P_\alpha{}^2=P_{A'}{}^2=0$.

A longitudinal polarization detector gathers *all* the longitudinal polarization information from a beam, whereas a transverse detector leaves unexplored the complementary component of transverse polarization. It is possible however to arrange to collect concurrently all the transverse polarization information in a beam of particles or photons, in which case the detector is not characterized by a single angle ϕ (supposing α to correspond to the beam direction) but strictly consists of two or more detectors operating in parallel.

C. Right *versus* Left Helicity

A right-handed particle or photon means one that "rotates" as it moves in the same manner as a standard right-hand screw. This can also be stated: When angular momentum and velocity are parallel the particle is right-handed, when antiparallel, left-handed.

In classical optics the opposite definition is in use: Right circularly polarized light is that in which the electromagnetic vectors show clockwise rotation as viewed by an observer into whose eyes the light enters, that is, an observer facing *opposite* to the direction of propagation of the light.

III. EXPERIMENTAL METHODS

A. Classification of Methods

Three different ways of analyzing spin polarization have been experimentally proved. One method makes use of the spin-orbit interaction when the particle is scattered by the Coulomb field of a nucleus, the method proposed originally by Mott (M29). Here any transverse polarization generally leads to a right-left asymmetry in the scattered intensity. This amounts to talking about a "force" $-\nabla(\boldsymbol{\sigma}\cdot\mathbf{H}')$ for e^-, where $\mathbf{H}'$ is the magnetic field experienced by the particle in moving through the Coulomb field $\mathbf{E}$, namely $-(\mathbf{v}/c)\times\mathbf{E}$. In the second method one makes use of the $\boldsymbol{\sigma}\cdot\mathbf{H}$ energy, where $\mathbf{H}$ now stands for an applied magnetic field. This method finds application where positronium is formed from polarized positrons. The third method is where one *compares* the polarization of the particle to be analyzed with a *reference* polarization residing in atomic electrons which are already aligned. In this category belongs use of electron-electron scattering (and positron-electron scattering), annihilation of very low energy positrons in a ferromagnetic material, or use of annihilation-in-flight of more energetic positrons in a magnetized foil. Also there belong here the second-order techniques where the particle polarization information is transferred to helicity of a photon, this helicity being then analyzed by Compton scattering against the polarized reference electrons.

B. Mott Scattering (and Spin Precessors)

Properly, the Mott scattering method applies to transverse polarization analysis only. For reasons of Coulomb repulsion it is not well suited for positron analysis. As one suspects from contemplating the fingers, an "up" spinning electron approaching "straight ahead" a nucleus of (positive) charge Ze should prefer to scatter to the right rather than to the left because of the force $\nabla\boldsymbol{\sigma}\cdot(\mathbf{v}\times\mathbf{E})$ implied by its magnetic moment moving

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

in the steeply graded field $\mathbf{B}'=\mathbf{E}\times\mathbf{v}$. The quantum mechanical results, first estimated by Mott and more recently computed by Sherman (S56), predict generally large right-left asymmetry in the scattered intensity. The azimuthal variation of the scattered intensity at angle ϕ is $\sin(\phi-\phi_0)$ for an incident electron polarized at azimuth ϕ_0. Conversely, the asymmetry in intensity between azimuth $\phi+90°$ and the conjugate azimuth $\phi-90°$, defined as

$$\delta_\theta(\phi)=\frac{I(\phi+90°)-I(\phi-90°)}{I(\phi+90°)+I(\phi-90°)} \tag{4}$$

is a measure of the transverse polarization at azimuth ϕ. A coefficient of proportionality relates the thing we presume to be measuring, the transverse polarization $P_s(\phi)$, and the observed δ. We call this coefficient S_θ following Sherman, and write $\delta_\theta(\phi)=S_\theta P_s(\phi)$.

For given energy, there is a best region in polar scattering angle θ at which to work, from the point of view of size of the asymmetry S_θ. Some of Sherman's results for the right-left asymmetry for an unscreened point scatterer ($Z=80$) are given in Table I. This shows at each of several speeds v/c, the maximum S_θ, the angle $\theta_{\max}$ at which it occurs, and the two adjacent angles called $\theta_{\frac{1}{2}}$ at which the asymmetry is down to one-half maximum value. At a certain angle θ ($\sim 60°$ at $v/c=0.4$ and shifting to $\sim 30°$ at $v/c=0.9$) the asymmetry S_θ becomes zero and changes sign. To avoid ambiguous factors of two, note that $S=+0.4$, for example, means that the intensity to the right (see beginning of this subsection for orientation) exceeds that to the left for an up-spinning electron incident, the ratio of intensities being 1.4/0.6.

What is not shown in Table I, but which could be very important when considering possibilities of systematic error in an actual experiment, is the variation with θ of the ordinary differential cross section. This gradient turns out to be quite steep at angles of experimental interest. A representative value, for the speeds shown in Table I, is 150% change in differential cross section per radian at $\theta=90°$. The gradient at $\theta_{\max}$ ranges from roughly 70% per radian for $v/c=0.4$ to about 150% per radian for $v/c=0.9$.

Also important is the fractional change in ordinary differential cross section for given fractional change in energy. At $\theta=90°$ this amounts to about 1.6, and would be somewhat larger if put in terms of fractional change in momentum.

TABLE I.

v/c	$S_{\max}$	$\theta_{\max}$	$\theta_{\frac{1}{2}}$	
0.9	0.51	150°	101°	170°
0.8	0.48	137°	90°	167°
0.7	0.45	130°	85°	165°
0.6	0.43	126°	83°	162°
0.5	0.40	123°	82°	161°
0.4	0.37	121°	85°	159°

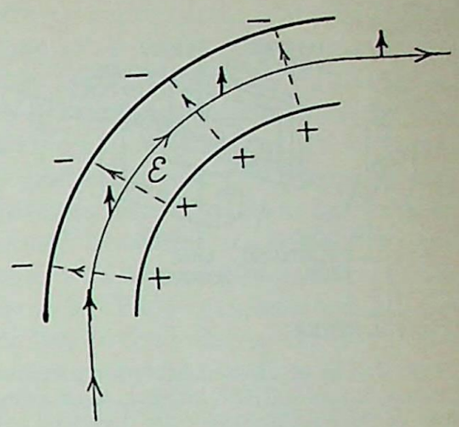


FIG. 1. Electrostatic spin-precessor due to Tolhoek and DeGroot (T51).

In the first successful detection of longitudinal polarization of beta particles from unpolarized nuclei by Frauenfelder and co-workers (Fa57), a means of first precessing the spin from longitudinal to transverse had to be employed in order to be in a position to use Mott scattering as the detector. A device originally proposed by Tolhoek and DeGroot (T51) was employed in order to "erect" the spins. Such a spin-precessor is shown in Fig. 1. It is essentially a nonrelativistic device and so the energies at which it works best are well matched to a Mott-scattering analyzer. Neglecting relativistic considerations, its action is simple and direct: The spin has no way to know that the electric field is continually changing the particle momentum $\mathbf{p}$, and thus the spin tends to retain its original direction in space, independent of sign or magnitude of the g value. (The device could more accurately be described as a momentum precessor.)

In the Frauenfelder experiment the Coulomb scattering act, (taking place in a thin gold foil) in analyzing the transverse polarization as the beam emerged from the spin erector, was measuring the longitudinal polarization that had entered the spin erector. Inasmuch as this spin precessor is a reversible device, that is, no information is lost through its use, we could classify this experiment as first order.

The angular aperture used was 95°–140°. A pair of end-window Geiger counters recorded the right- and left-scattered betas. The scattering foil was either 0.05 or 0.15 mg/cm² gold on a low-Z backing. The results of the first measurement, using a Co^{60} source, are well known: These negative betas were predominantly left-handed, the magnitude of the asymmetry obtained being taken to imply longitudinal polarization at emission of magnitude v/c, at least at sufficiently high beam energy that all the "dirt" effects that had at times plagued certain of the double-scattering experiments were probably not serious. There was a drop in apparent polarization of over a factor two on changing the mean v/c from 0.49 to 0.47, the decrease amounting to essentially two standard deviations.

We continue to introduce Coulomb scattering experiments in terms of the kind of spin precessor used, because the Mott scattering approach to polarization is

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

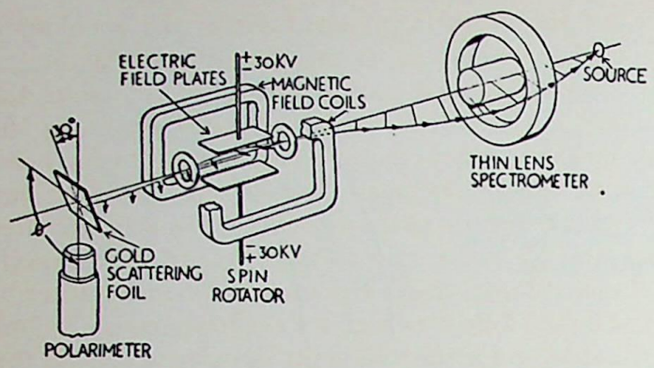


FIG. 2. Use of crossed-field spin-precessor. In the apparatus of Cavanagh *et al.* (CR57) original longitudinal spins entering at the right of the spin-precessor are precessed to emerge pointing downward. Polarimeter unit consists of scattering foil plus scintillation counter which can rotate about the beam axis.

scarcely new in the laboratory, and has been reviewed (T). Actual use of spin precessors is however quite new.

Another spin precessor is the crossed-field device. The beam is arranged to travel at right angles to an electric field **E** and a magnetic field **B** which are themselves at right angles. In the type to be described the fields are homogeneous over a certain region (it may be clearer to say that the geometry is rectangular). The crossed fields define a speed $\beta_0 = E/B$. If one arranges $0 < \beta_0 < 1$, then in the moving system characterized by speed β_0 in the laboratory, there is no electric field, and the magnetic field is $B' = B(1-\beta_0^2)^{\frac{1}{2}}$. A Dirac particle at rest (for example) in this moving system has its spin precessed at angular speed $\omega' = eB'/m_0c$, the sense depending on the sign of the g value. In the laboratory system this means that a particle that has traveled a distance $z = 2\pi\rho_0(1-\beta_0^2)^{-\frac{1}{2}}$ has its spin precessed through an angle $\theta = 2\pi$. Here, $B\rho_0$ stands for laboratory momentum of the central ray. For distances in the laboratory Δz, differing from integral multiples of the distance just cited, the precession angle of a general ray may deviate somewhat from that of the central ray, which remains $\Delta\theta = (1-\beta_0^2)^{\frac{1}{2}}\Delta z/\rho_0$. But such deviations are readily computed.

A crossed-field device with suitable slit system can be a mass spectrometer,‡‡ and not simply a "velocity-filter." It might be possible to make practical use of this fact in dealing with polarized particle beams.

Cavanagh and co-workers (CR57) have used a crossed-field device as a spin-erector. The beta energy was preselected by means of a magnetic lens (Fig. 2). This lens does nothing to alter P_p since $\boldsymbol{\sigma}\cdot\mathbf{v}$ in a region where $\mathbf{E}=0$ is a constant of the motion. Next the crossed-field precessor interchanges longitudinal and transverse polarization components in the beam. Finally, Mott scattering at nominal $\theta = 90°$ is used to analyze for transverse polarization. The scintillation detector (of size characterized by 0.3 radian) plus the thin gold scattering foil form a unit, the polarimeter unit, which may be rotated in azimuth ϕ. The results show the expected right-left asymmetry [the observed δ as in Eq. (1) being as large as 20%], but in addition a curious and hard to explain up-down asymmetry presumably of instrumental origin. It was possible by reversing fields to split off this latter (orthogonal) asymmetry from the desired one. In the first experiments Co^{60} and Au^{198} negative betas were shown to have longitudinal polarization consistent with $-v/c$. Alikhanov *et al.* have also used a crossed-field spin precessor (A58).

Regarding attenuation of the sought-for asymmetry due to the scattering foil having finite thickness, Fig. 3 shows an extrapolation against foil thickness. This gives a graphic demonstration of how finite foil thickness, presumably through plural scattering and possibly depolarization, reduces the observed asymmetry.

A third device for transferring polarization information from longitudinal to transverse (or vice versa) has been used specifically to prepare a beta beam for Mott scattering analysis. This we call a "spin reprojector." It consists in scattering the original beam by atoms, whether singly or multiply does not matter to a first approximation. The transverse polarization of the electrons after scattering at $\theta = 90°$ represents the input longitudinal polarization, to the extent that the original spin direction is unchanged on scattering.§§ The action is similar to that of the macroscopic electrostatic spin precessor.

The resulting transverse polarization lies in the plane of the scattering. Even taking into account the fact that the spin-orbit effect may not be entirely negligible, this cannot lead to any polarization in this plane. The spin-orbit effect and any helicity memory act only in competition to dilute and dull somewhat the transfer efficiency of interest, since the sum of the squares of

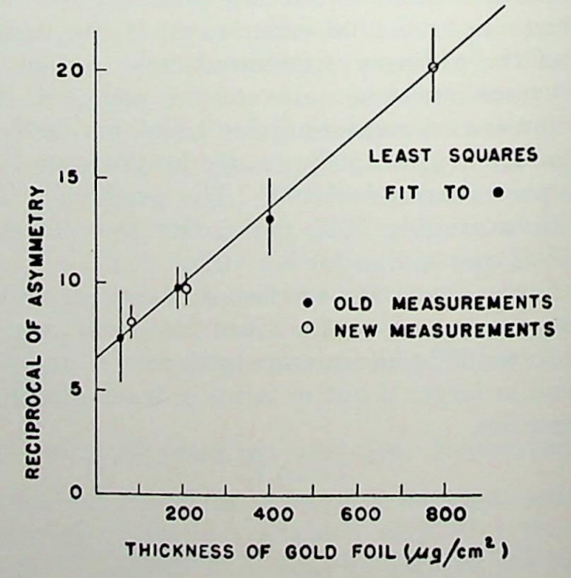


FIG. 3. Effect of scattering foil thickness on the measured asymmetry, from Cavanagh *et al.* (CR57).

‡‡ Compare the nonrelativistic applications by Bleakney and Hipple (B38).

§§ Bernardini, *et al.* (B58) treat this matter quantitatively. They show that there can be considerable helicity memory in a Coulomb scattering event.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

three mutually complementary polarizations cannot exceed unity. A fairly low-Z scattering material is preferable from this point of view.

Spin erection by means of this reprojection method has been used by Pond (Po57) on P^{32} electrons. It has been used in a very nice geometry by Deshalit and coworkers (De57). Their apparatus is shown in Fig 4. The spin reprojector is in the form of a sheet of aluminum (denoted by symbol σ_1) forming an arc of a circle. From the geometry of the circle one may then have (for point source and point second-scatterer) constant "total" scattering angle θ, in this case 90°, over a large range of emission angle at the source. This affords a gain in intensity which is needed in a second-order experiment. The resulting asymmetry for P^{32} electrons of energy exceeding 900 kev, at nominal second-scattering angle $\theta=75°$, was 5%. This was interpreted as compatible with full v/c polarization.

Spin erection by means of Coulomb scattering was suggested independently also by Tassie (Ta57). Another possibility for spin erection is based on Møller scattering [see Sec. III D (1)].

A high-energy precession effect relating to the anomalous magnetic moment has been computed by Mendlowitz and Case (M55). A spin precessor based on such an effect, used in conjunction with a Mott scattering analyzer might suffer from a limitation due to energy mismatch. To illustrate with a simple situation where a Dirac particle travels normal to a magnetic field, if the g value were simply two, this would cause the preservation of spin polarization projected onto the velocity vector. In terms of "time dilation," the zero-energy value of the cyclotron frequency (the frequency of rotation of the velocity vector) and the spin precession frequency are each multiplied by the same factor, $(1-\beta^2)^{\frac{1}{2}}$. The two quantities, velocity and spin direction, then stay in phase for a hypothetical Dirac particle. Otherwise said, $\boldsymbol{\sigma}\cdot\mathbf{v}=\boldsymbol{\sigma}\cdot(\mathbf{p}-e\mathbf{A})$ is a constant of the motion because it commutes with the Dirac Hamiltonian. The interesting result of the Mendlowitz and Case calculation is that the additive "Schwinger part" of the magnetic moment (or of the g value) is predicted to be immune to this time-dilation factor.

If the velocity of the particle is effectively along the magnetic field (so that we might transform along the field, find the particle moving only normal to the field, but with practically nonrelativistic energy) then the spin ought to precess at the rate of $\omega_{cyc}(1+\alpha/2\pi)$. Higher terms are not mentioned because free-particle experiments are still concerned with the first||| non-Dirac term. It would then take the order of 137 cyclotron revolutions for the spin vector relative to the velocity vector (in the moving system) to change by one radian. However, for the case of primary interest here, where the particle with substantial energy moves normal to the magnetic field, it would require *only* about $137/E$ cyclotron revolutions to precess the spin one radian further than the velocity vector. Here E is the total energy of the particle in units of its rest energy.

This computation has recently been done again by Carrassi (C58) using a somewhat different approach, with the same result obtained by Mendlowitz and Case. Such an effect would be of interest for itself, independent of any "practical" use. For de-erecting transverse polarization of highly relativistic particles it could possibly be useful, since P_s becomes increasingly difficult to measure directly as kinetic energy increases above the rest energy.

To return to the Mott scattering method itself. It would seem that the effect of the screening of the nuclear charge by the atomic electrons will have to be understood quantitatively, as will plural and multiple scattering effects in finite scattering foils, before precision polarization numbers can be obtained with confidence. These matters are under active scrutiny in a number of laboratories, and therefore after a time a good understanding of this useful method must emerge.¶¶

C. Energy Term $\boldsymbol{\sigma}\cdot\mathbf{H}$

The Pauli energy term $\boldsymbol{\sigma}\cdot\mathbf{H}(e\hbar/2mc)$ affords a means of determining whether the spin $\boldsymbol{\sigma}$ is parallel or antiparallel to an applied magnetic field $\mathbf{H}$. For measuring σ_z of free beta particles it has so far not been employed.*** Niels Bohr showed that a Stern-Gerlach experiment on bare electrons is not a good way to go about such a measurement. This has been discussed thoroughly by Mott and Massey (MM).

One may however capture the particle of interest into a bound system (see again the discussion by Mott and Massey) and then measure $\boldsymbol{\sigma}\cdot\mathbf{H}$. Concerning the preservation of the spin direction during capture, one notices that it is the charge e of the particle and not its

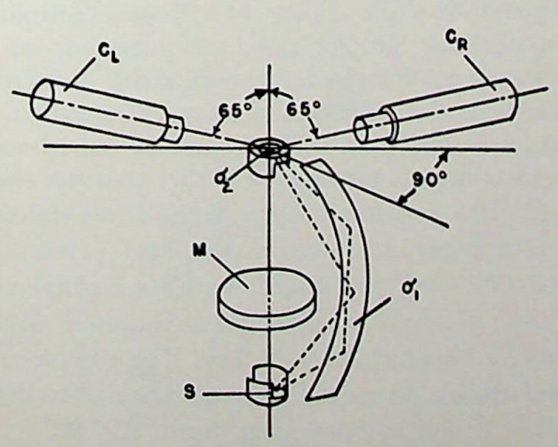


FIG. 4. Apparatus of DeShalit *et al.* (De57), illustrating "spin reprojection" by a first scattering of the betas from source S in the curved aluminum sheet, σ_1. Analysis of transverse polarization lying in the plane of $S-\sigma_1-\sigma_2$ is effected by Mott scattering at σ_2.

||| *Note added in proof.*—However, see recent precision results for g factor of a free electron obtained by the University of Michigan group, Bull. Am. Phys. Soc. Ser. II, 4, 250 (1959).

¶¶ It will then (by definition) be possible to perform a double-scattering experiment to high precision, with no anomalies.

*** A resonance experiment on free electrons has been performed by Dehmelt (D58).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

magnetic moment that makes the process go. Hence, so long as the energy transfers involved (usually tens of volts only) are small with respect to m_0c^2 of the particle, there is no spin-flipping *mechanism* during the act of capture.

Let us specialize to formation of ground-state positronium since this has been employed for polarization measurement. Here an "up" positron allowed to form positronium (say by radiative capture with an effectively free electron) will not leave electron holes more "up" than "down." The fine structure of the resulting positronium is too small to cause selective capture to singlet rather than to triplet states (the former being bound more tightly by ΔE of only about 10^{-3} ev). It is necessary to examine a little more closely to find out where the original positron polarization information goes.

Consider still a free "up" polarized positron (that is with its spin along $+z$) which eventually is captured by one of a group of electrons which as a whole had no spin preference with respect to the z direction. Suppose there are no applied fields. On a time scale, the triplet-singlet energy difference ΔE is by no means inconsequential: ΔE corresponds to $\sim 10^{-12}$ sec. Thus at a time later than capture of the order of 10^{-12} sec, a (resonably adiabatic) inquiry whether the positronium state is singlet or triplet makes physical sense. This arises spontaneously in the process of decay by annihilation into two or three photons, respectively. At this stage, having allowed singlet and triplet state amplitudes to have lost their relative phase, the original spin information is largely lost: To be sure it could be partially retrieved if one looked only at certain kinematically special triplet (three-quantum) annihilation events, measuring photon helicity—rather difficult experimentally. For singlet annihilation (into two photons) there is no information that can be traced back to the original positron polarization, either in the angular correlation with respect to the $z=0$ plane, or in the helicities of the photons, the parent state being $J=0$ [see, for example, Yang's paper on dematerialization, (Y50)]. In the case of the singlet state, the original polarization has not actually been lost, but resides in the sign of the hole where the captured electron was, with complete preservation of information. For the three triplet substates as a whole the information resides 67% in the helicity of the residual photons, and 33% in the sign of the hole.††† The important point is that by examining only the two-photon residue, no information remains. This is in contrast to higher energy direct annihilation where quite large retention of polarization information is possible [see Sec. D (5)].

If the positronium is formed in the presence of an applied magnetic field, along z, say, then in effect some phase coherence between singlet and triplet states (of same J_z, thus only $M=0$ states are affected) can be "frozen in" until the annihilation takes place, and not only for the 10^{-12} sec permitted by the triplet-singlet energy difference. Under these circumstances some polarization information will now be retained. We should recall how the Zeeman splitting goes for ground-state positronium [see the positronium review articles (D53) and (D54)]. Here, J is no longer a good quantum number in the presence of the field $\mathbf{H}$, but only its projection along the field. Let us use an arrow for up or down positron spin, and likewise for the electron spin. Let the positron's spin state stand first. The state $\uparrow\uparrow$ and the state $\downarrow\downarrow$ are each unperturbed by the applied field. The $M=0$ part of the Hamiltonian, normally diagonal, now has, in virtue of the $\boldsymbol{\sigma}\cdot\mathbf{H}$ energy, part of which belongs to the positron and part to the electron, matrix elements connecting triplet and singlet. On rediagonalizing the Hamiltonian, the *nominal* triplet state is $(\uparrow\downarrow+\downarrow\uparrow)-\epsilon(\uparrow\downarrow-\downarrow\uparrow)$, and the nominal singlet state is $\epsilon(\uparrow\downarrow+\downarrow\uparrow)+(\uparrow\downarrow-\downarrow\uparrow)$. The mixing amplitude ϵ is practically linear with H, below 20 kilogauss, and becomes unity at extremely high field. The signs are such that positive H_z means positive ϵ.

In the high field limit, the two states of interest become (for positive H_z) $\downarrow\uparrow$ for triplet and $\uparrow\downarrow$ for singlet, indicating complete spin memory. Even at modest field values, say about 15 kilogauss, the disparity in $\langle\sigma_z\rangle$ for the positron (or the electron) between triplet and singlet is about a factor two. We can think of this as a capture efficiency $\epsilon_{cap}=-0.33$. This means that initial positron polarization (called P_z) would be 33% preserved in the capture act, residing finally in the asymmetry between the population of triplet and singlet states. As explained earlier, capture to $M=1$ states becomes generally a three-quantum annihilation problem and is not considered.

The discussion now divides according to whether the positronium is formed in a gas or in condensed material.

Positronium in gases.—At a field strength of 15 kilogauss about 0.97 of the triplet ($M=0$) states annihilate via two photons, and for practical purposes, *all* the singlet states do. Except for a peculiarity relating to the kinetics of the positronium atom just formed, the polarization information (all but some 3% of it) would now be lost, since it was agreed earlier that we should not attempt to measure helicity of the three quantum events. The triplet state has at this field value a lifetime $\sim 3\times 10^{-9}$ sec, the singlet state its normal value $\sim 10^{-10}$ sec. Nascent positronium in argon is known to have recoil energies of 1 or 2 ev. The magnetically quenched triplet states yield two photons observed to have such strong angular correlation at 180°, when suitable slowing (thermalizing) agents are present in the gas, that undoubtedly the positronium has lost essentially all its original recoil energy in 3×10^{-9} sec.

††† Thus two-thirds of the time, J, $M_J=1$, $+1$, and one-third of the time 1, 0, restricting to final $J=1$. The transfer efficiency from initial positron polarization to hole polarization is thus one-third.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

This behavior was known some time before polarization became of interest (H56, He57).

One may judge from the observed mutual angular correlation for the two-photon events whether the parent state was nominal triplet or singlet, the connection between an observable and the original positron spin direction being thereby completed. With reasonably narrow slits (about 1×10^{-3} radian of projected angle) the former states can be counted about five times as efficiently as the latter, at first glance implying the large efficiency for this stage of the operations of 0.67. However, inasmuch as there are many competing annihilations recorded (about five times as many positrons as we have just discussed are at the same time making competing direct annihilations in the gas sample) the effective efficiency because of the diluting action of such events drops to about 0.20.

To summarize: Let a hemisphere of positive betas emitted in the general z direction come to rest in a gaseous annihilator without depolarization. The geometrical efficiency (for projecting any longitudinal polarization P_p onto the z direction, calling the latter polarization P_z) is then 0.5. The capture efficiency ϵ_{cap} to ground-state positronium as discussed earlier would be 0.3 (at 15 kilogauss), the detector efficiency as discussed immediately above could be typically 0.2. The expected δ (on reversing **H**) is then $2\times0.5\times0.3\times0.2\times P_p = 0.06P_p$. In the initial experiment of Page and Heinberg (PH57) it was found that Na^{22} positrons characterized by v/c at emission of 0.75 were polarized with $P_p > +0.30$, the lower limit being considered conservative in that no backscattering corrections from the gold backing nor depolarization estimates were made (Fig. 5). The detector efficiency was purposely overestimated.

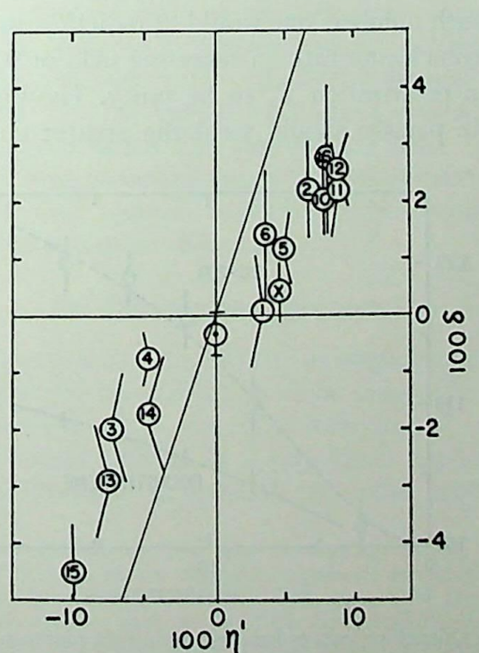


FIG. 5. Use of positronium in a gas as polarization detector (PH57). For longitudinal positron polarization at emission of +0.75, measured δ would follow oblique line whose slope is proportional to polarization.

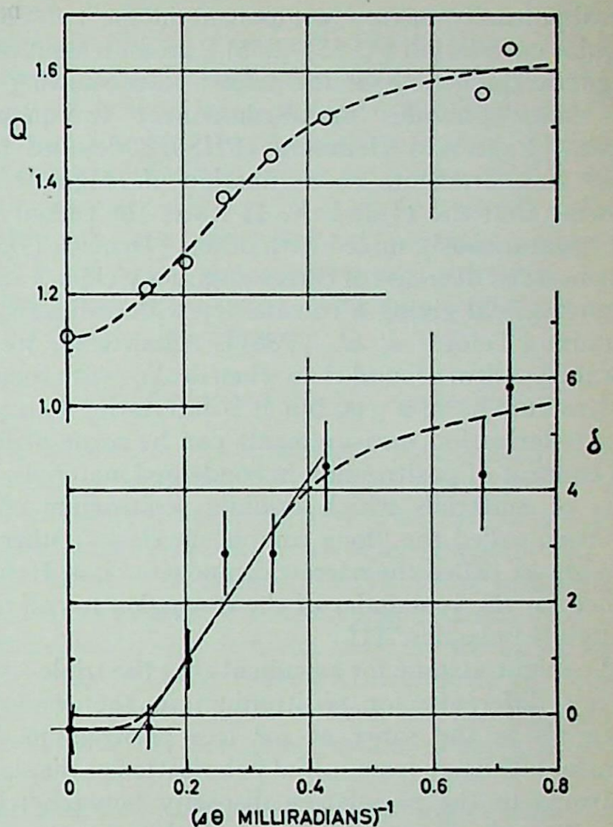


FIG. 6. Positronium in a gas (PO58). Lower curve: extrapolation of measured δ to zero effectively yields a measure of width of singlet contribution to the two-photon angular correlation. Upper curve: quenching ratio Q at infinite slit width yields fractional positronium formation in gas sample. Variation of Q with slit width yields width of triplet events.

Later data (Fig. 6), where effectively this latter efficiency is measured for the particular gas sample in use, by extrapolating measured δ against reciprocal angular resolution, gave $P_z=0.25$. This implies a lower limit on P_p of +0.50. The low-Z source-backing in the latter experiment (quartz) improved matters somewhat. Detailed angular correlation measurements (PO58) showed clearly the transfer from "broad component" to "narrow component" when the H field was changed from parallel to the positron momentum to antiparallel.

In a sense this kind of experiment is a comparison one: between the sign of $\langle\boldsymbol{\sigma}\rangle\cdot\mathbf{H}$ and the theoretical sign of the triplet-singlet energy difference for ground-state positronium. Here it was of practical concern for the first time which lay higher in energy, the triplet or singlet states.

Positronium in condensed materials.—Since the time when DeBenedetti and Richings (D52) found positrons to live longer in certain dense materials (specific gravity unity or more) than free positrons should, a number of diverse experimental results have been in accord with the hypothesis that positronium may exist in a certain class of substances. We mention the detailed measurement of these long lifetimes (Bell and Graham B53), the measurement of the abnormally high three-quantum yield [Pond (P54), Graham and Stewart (G54)], of

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

the abnormally narrow component in the two-photon angular correlation [(P55) (S55)], measurement of the magnetic quenching of the triplet state showing that this state provides broad-component two-quantum events [Page and Heinberg (PH56)], detailed time-delay measurements as a function of magnetic field showing that the ↑↑ and the ↓↓ states (in Teflon) were not spontaneously mixed with others [Deutsch (D56)], the observed decrease of three-quantum yield in a strong magnetic field giving a consistency with ordinary positronium [Telegdi *et al.* (T56)]. Admittedly, we are now on less firm ground than when dealing with *bona fide* positronium as in a gas, but it is interesting that positron polarization measurements can be made utilizing the concept of positronium in condensed materials. The class of materials which exhibits positronium effects has been called the "long component class." Otherwise it might be called the narrow component class. It might suffice for the remainder of the discussion to call these materials "plastics."‡‡‡

We might assume for argument that the triple-singlet energy difference for positronium in the condensed materials is the same as for free positronium. The transfer efficiency from initial polarization of the slowed positrons to the population disparity between triplet ($M=0$) and singlet states would then be the same figure at the same magnetic field as for free positronium. From this point on, the picture is quite different.

For our purposes, each positron which comes to rest in the annihilator gives two photons. Ignoring polarization for the moment, let us discuss a typical "plastic" such as fused quartz. The original conjecture was that the narrow component occurred in parallel§§§ with the delayed component (P55). If we retain this view, and also pretend that all measurement is ideally precise, we shall have the singlet states giving only narrow component, denoted N, the triplet states in absence of applied field giving only broad component, denoted B, with nonpositronium events giving also a broad angular distribution, B'. In actual measurement, B' has never been distinguished from B. Discussion is now restricted to the two $M=0$ states, singlet and triplet, unless otherwise stated.

Application of a strong field H, strong enough to compete successfully with the rather fast (one or two by 10^{-9} sec lifetime) "pickoff annihilation," causes a detailed transfer of events from B to N [(Pa55) (PH56) (W57)]. Fields of 10 to 20 kilogauss are typically required. If an angular correlation apparatus is set to accept only 180° events, within about 10^{-3} radian, the coincidence rate increases with H, as in the so-called quenching curves of Fig. 7. Here the detector is perhaps four times as efficient for N events as for B events.

Regarding the possibility of positron polarization measurement, the situation in one respect is not as fortunate as with positronium in gases: "up" positrons in an "up" field ought to form more singlet states than triplet, the singlet states give only N events, but the triplet states instead of giving only B events (as at zero applied field) now might divide 0.6 of the time into B and 0.4 of the time into N, for example. In other words the H field which is required to furnish the initial "capture efficiency" acts unfortunately to dull the action of the detector from distinguishing triplet from singlet events.

In order to see what response to expect for plastics let us supply some rather rough numbers, and assume that the field is such that one-third of the triplet ($M=0$) decays appear as N events. For orientation, at $H=0$, one has 0.5 of the positrons of interest forming singlet states and 0.5 triplet ($M=0$) states. The fraction f_N of these yielding N events is thus 0.50. For unpolarized input positrons, at $H=H_{\frac{1}{2}}$, one has still the equal division into singlet *versus* triplet states, but f_N becomes now 0.67. For completely polarized stopped positrons, at $H=H_{\frac{1}{2}}$, the capture efficiency for keeping the polarization information might be 0.33 (depending indirectly on the lifetime of the delayed component). Thus, for field **H** parallel to the spins, a singlet/triplet population ratio of 0.67/0.33 is implied, and f_N is computed as 0.78. For the opposite field setting one gets $f_N=0.55$. In order to arrive at a final number, suppose that 0.35 of *all* incident positrons go into ($M=0$) positronium, and that the detector response $N:B$ is 4:1. Using such numbers one would "predict" a measured δ in the coincidence rate on reversing field of 0.14, given positron polarization P_z to be unity. Having field **H** and spin parallel should yield the greater coincidence

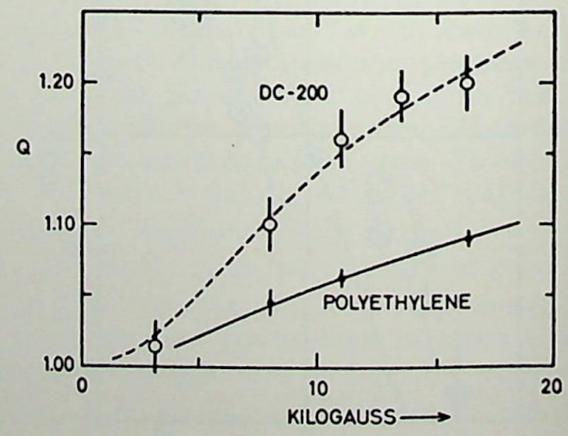


FIG. 7. Quenching curves for positronium in plastics. Fractional increase in two-photon coincidence rate within 1.5 milliradians of 180°, as a function of applied field H, corrected for focusing effects. Actual transfer from broad B component to narrow N component on applying field is less than one-fourth of $Q-1$. From Page and Heinberg (PH56).

‡‡‡ Inasmuch as whether or not the material tends to harbor positronium seems to connect with whether or not it is crystalline, some one has facetiously suggested the following nomenclature: materials not having positronium effects shall be Xtals, materials having them shall be Ytals. (This would save the term Xtal from extinction in the physical journals.)

§§§ Thus a positronium annihilation event is either narrow or it is delayed (at $H=0$). Attributing the delayed events to triplet positronium, one expects from statistical weights three times as many delayed events as narrow events.

CC-0. Central Kangri University Haridwar Collection. Digitized by eGangotri Foundation USA

rate, which is opposite to the response in the gas experiments just described.

The magnitude of the response of plastics to polarized positrons is next to impossible to estimate *a priori* because so many essentially unknown numerical factors enter. But the *sign* of the response may be traced back to one hypothesis, that the normal fate of the triplet states is broad annihilation (a half-width like 10×10^{-3} radian), and the normal fate of the singlet states is narrow annihilation (half-width less than 3×10^{-3} radian).

To come finally to an actual experiment, Page and Obenshain (PO57) used a hemisphere of Na^{22} positrons (gold-backed) which impinged on the plastic sample. The δ in the coincidence rate at 180° with angular resolution 1.5×10^{-3} radian was measured on reversing **H**. Only for the narrow-component materials was a nonzero δ found, and its sign agree always with that anticipated in the foregoing discussion. With no absorber between source and sample, and running typically at $H=13$ kilogauss, δ was measured for fused quartz, polystyrene, DC-200 high viscosity silicone oil, polyethylene, and Teflon. Expressed in percent the results in order were 1.7 ± 0.5, 1.7 ± 0.7, 1.3 ± 0.6, 1.3 ± 0.4, and 1.9 ± 0.5. With an interposed absorber of 0.0005-in. tantalum, the Teflon response increased to $4.8\pm1.1\%$, while a (low Z) Mylar foil of like density in mg/cm² increased the Teflon response to $4.5\pm0.9\%$. The very roughly estimated response quoted above would yield 7% expected δ for a hemisphere of fully (longitudinally) polarized positrons, neglecting back reflection.

Hanna and Preston (Ha58) have checked the sign of the response of polyethylene against the "iron detector." Page and Obenshain have checked the sign of the iron detector against the plastics [see Sec. III D (2)].

The fact that plastics in general give a detectible response to positron polarization must be taken as evidence that depolarization on slowing is small (even in tantalum, $Z=73$) and as supporting evidence that something very much like positronium must exist in this class of materials, and finally that the separation‖‖‖‖ into broad and narrow components for triplet and singlet states is rather clear-cut.

D. Comparison Methods

As discussed earlier (Sec. III A) a comparison method is one where the spin direction to be analyzed is compared somehow with the spin direction of atomic electrons already known. It is sometimes implied that comparison methods in general are rather inefficient. However, if one is discussing the particular physical act which is the heart of a given method, then it often turns out that efficiencies are closer to unity than to zero. It is best to keep separate the physics of the analysis act from the question of what degree of (reference) polarization might be achieved in practice for the average electron in a typical "iron" foil for example. In the discussion of several comparison methods, efficiency means efficiency per unit degree of polarization of the target electron, unless some statement is made to the contrary.

(1) *Møller and Bhabha Scattering*

The most direct and intimate comparison of spin directions is afforded by the electron-electron and positron-electron scattering processes, which we label "Møller" and "Bhabha" scattering respectively. The matrix element of Møller (M32), from which have stemmed the well-known Møller and Bhabha scattering formulas, has within it much polarization information which, prior to about 1957, was not much considered. Bincer (Bi57) seems to have been the first to publish computations on the polarization aspects of such scattering.

Speaking first about electron-electron collisions, and in the nonrelativistic limit, elementary quantum mechanics tells us that if a pair of identical spin-½ fermions are to undergo mutual scattering (letting all spins be random), one-half of the time the spins are antiparallel, the particles therefore distinguishable after the collision (they are always distinguishable before the collision) and the scattering cross section has its classical value at all angles in that there are now no interference (exchange) effects. For the other half of the time the spins are parallel, in which case we have indistinguishability after the collision, interference effects come into play, and, for 90° scattering in the center-of-mass system in particular, there can be no scattering. On the whole there is a progressive shift from the classical cross section at small angles to one-half of the classical value in the large angle limit, 90°.

The alternative view (which is not particularly useful for polarization discussions) is that three-fourths of the time the pair of spins form a triplet state, nullifying the scattering at 90°, while the remaining one-fourth of the time they are in a singlet state, now with constructive interference for fermions which doubles the cross section at 90°. The same numerical result is obtained at any angle as with the previous view. This result hinges on there being no spin-flips during the collision and again no spin sensitive forces in the scattering "force," which guarantees the first proviso.

Longitudinal polarization analysis.—For the relativistic discussion, let us focus attention on helicities. Writing the appropriate free particle spinors for definite initial and final helicities, then applying the Møller interaction and subtracting Feynman diagrams,

‖‖‖‖ Further results by Bell and co-workers (GB57) on the fraction of positrons in the delayed component for fused quartz in particular (now reported to be 0.51 instead of the earlier estimate of about 0.3) have removed the numerical disparity that existed for some time between fractional narrow component and fractional delayed component. The recent polarization applications however, being based on a balance method, might be argued to lend strong support to the contention that this separation is clear-cut.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

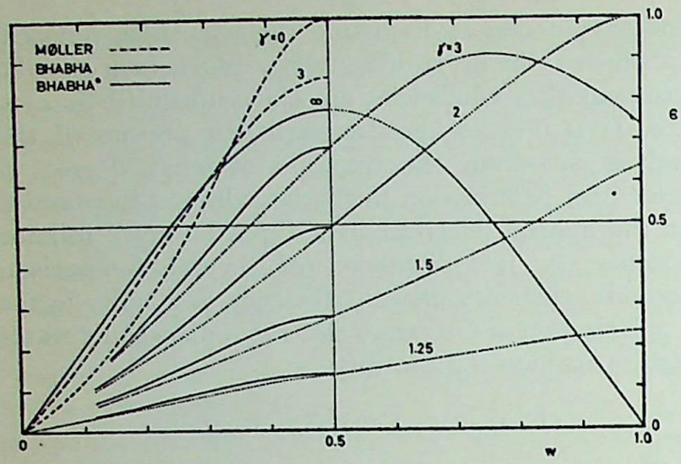


FIG. 8. Differential Møller and Bhabha scattering. For a completely polarized target electron, the efficiency ϵ for detecting longitudinal polarization of e^- or e^+ respectively, is plotted against fractional energy transfer w. Total energy of incident particle is γm_0c^2. Møller and Bhabha expressions coalesce as $\gamma \to \infty$. Solid curves, labeled Bhabha,* apply if final e^+ and e^- are *not* distinguished by the apparatus. Inasmuch as cross section generally changes rapidly with w, partial areas under curves may have but limited practical significance. Curves have been plotted from the formulas of Bincer (Bi57).

that is, writing down the Møller matrix element, the following facts emerge. We refer mostly to 90° scattering in the center-of-mass system where those polarization effects which arise entirely from the subtraction of diagrams are generally a maximum.

In the first place, single flips of helicity are taken out by the interference (at 90° scattering angle). In the second place, particle helicity tends to be better conserved as the energy increases.¶¶¶ We are not referring to the more trivial effect where at small angle the helicity naturally is well preserved. The helicity of one of the input particles would be in fact 80% conserved at only 0.6-Mev laboratory energy, in the sense that either of the output particles viewed in the center-of-mass system preserves 0.8 of the input helicity (for 90° scattering). This fact is used later. In the third place, and this is the important point upon which the polarization experiments have been based, the relative cross section (at 90°) between input spins parallel and antiparallel, for *longitudinal* polarization only, which was zero in the low energy limit climbs only to $\frac{1}{8}$ in the high-energy limit, the corresponding efficiency being therefore 7/9. The complete expression for this cross-section ratio as given by Bincer (Bi57) is

$$\frac{\sigma_p}{\sigma_a}=\frac{2x^2+\beta^2(3x^2+x^4)+\beta^4(1+x^2)}{1+x^2+\beta^2(2+3x^2-x^4)+\beta^4(5-4x^2+x^4)}, \quad (4)$$

where β is the speed of either particle in the center-of-mass system and $x=\cos\theta$ where θ is the center-of-mass angle of scattering. This result is shown in Fig. 8, now put in terms of analyzing efficiency ϵ for a longitudinally polarized target electron. The abscissa is the conventional fractional energy transfer, $w=\frac{1}{2}(1-\cos\theta)$. The two dashed curves labeled "Møller" and the limiting shape at high energy labeled ∞ apply here.

Bincer has also computed the companion expression for Bhabha scattering (Bi57). A few curves computed from these results are also shown in Fig. 8. These are the dotted curves labeled "Bhabha." At first glance, the e^+ and the e^- are distinguishable and thus one expects no polarization sensitivity in the sense of a correlation between positron and electron helicities for a given probability of scattering (of course, both initial and final e^+ helicity tend to be correlated, and the same applies for the e^-, on any simple theory). The point here is that there exists the virtual annihilation process, and this accounts for a good part of the ordinary (unpolarized) positron-electron scattering cross section (A54). This process is represented by one of the Feynman diagrams. It manifests itself now in making possible the polarization effect shown in Fig. 8. The effect dies out at quite low energy, where the annihilation amplitude has diminished.

Inasmuch as the Bhabha polarization efficiency ϵ as a function of w gives a rather unrealistic picture of how an experiment might be designed (because the differential cross section typically changes by a factor 30 between $w=0.25$ and $w=0.75$), we show also some modified Bhabha curves. These efficiency curves, labeled Bhabha,* would apply if the two particle detectors, operated in coincidence (or a single detector, if such an experiment were done), were completely insensitive to whether the counted particle is a final e^+ or a final e^-.

To summarize the picture presented in Fig. 8, the Møller method for longitudinal polarization measurement could be classed as a low-energy method, based on symmetry (Pauli principle)—the fact that a high efficiency persists to all energies can be regarded as good fortune. As for the Bhabha method, which again rests on symmetry, high energy is required to provide sufficient annihilation amplitude that the symmetry may come into play, so it is essentially a high-energy method. The left half of Fig. 8 clearly shows these energy trends. Another view is that at high energies the two methods are just the same (the infinity curve), but that at lower energies different details appear, simply because the forbidden zone of width $2m_0c^2$ is no longer negligible.

The Møller scattering method for the purpose of polarization detection was first employed by Frauenfelder and co-workers (Fb57) on negative betas from P^{32} and from Pr^{144} with the result that both beta groups were left-handed, and by an amount consistent with v/c at emission. Their apparatus is shown in Fig. 9. For reasons of impingement of beam, the scattering foil is inclined to the beam direction, so the efficiency has to be multiplied by the cosine of the small relevant angle. What is measured here is not an absolute differential cross section, or relative cross section against angle of

¶¶¶ This is not surprising since γ_5 does commute with the Møller interaction, but we remember that neither particle at finite energy can be in a pure eigenstate of γ_5.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

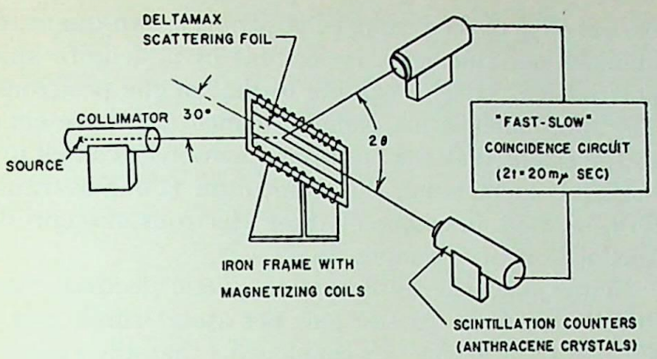


FIG. 9. Møller scattering setup of Frauenfelder and co-workers (Fb57).

collision, but only a comparison of a coincidence rate *proportional to* cross section between the two opposite states of magnetization of the scattering foil. For this reason the apparatus need not be especially elaborate.

The full definition of a true Møller event is that (i) $\mathbf{p}_1+\mathbf{p}_2=\mathbf{p}_0$. That is, momentum of particle No. 1 and particle No. 2 (labeled by which counter they enter) shall add up to the incident momentum. (ii) $K_1+K_2=K_0$. The kinetic energies likewise shall add up to the original kinetic energy. Coincidence detection of Møller events has been effected in the past by requiring for example that $(\mathbf{p}_1+\mathbf{p}_2)\cdot\mathbf{p}_0=p_0^2$ (P51), (A54), which is a compromise between complete rigidity of definition and counting rate. At energies of say 0.5 to 2.0 Mev, at the fairly high fractional energy transfers, say $0.4<w<0.5$, both $\mathbf{p}_1$ and $\mathbf{p}_2$ make sufficiently large angles with $\mathbf{p}_0$ that this simple criterion rules out multiple or plural scattering in the target foil, particularly after the Møller event, as well as against energy loss on the part of either particle No. 0, 1, or 2. Thus by setting (besides p_0) a certain value of $\mathbf{p}_1\cdot\mathbf{p}_0$, coincidences are found only at the expected value of $\mathbf{p}_2\cdot\mathbf{p}_0$ if the experiment is clean. See in this connection "coherence curves" (A54).

Returning to the experimental setup of Fig. 9, where a considerable spread was accepted in the incident energy K_0, typically a factor two or three, one would lose essentially all opportunity to check for "coherence" between K_1 and K_2. There does still remain the possibility for an "angular coherence" check, which however due to the relativistic narrowing is sharpest only for fixed K_1+K_2.

For the isotope Au^{198}, results obtained with the setup of Fig. 9 were interpreted (Fc57) as implying negligible longitudinal polarization for these particular negative betas (some seven standard deviations less than the expected v/c value). The Møller method was subsequently proven, for the case of Au^{198}, by Benczer-Koller *et al.* (BW58). The apparatus used by the latter group did not differ in essential details from that in Fig. 9. A magnetic lens system was used to bring the beam into the apparatus and focus it on the "iron" foil. It has been pointed out (W58) that precautions were taken in the latter experiment to discriminate against true time-coincident Compton electrons (due to the accompanying Au^{198} gammas) which could otherwise produce false Møller events, and that these precautions could probably account for the difference between the two results. The latter result was v/c polarization to seven standard deviations, and left-handed.

Use of the Bhabha method for longitudinal polarization measurement was first attempted by Frauenfelder and collaborators (Fd57) on the high-energy positron group from Ga^{66}. The results were interpreted as meaning that these positrons were probably not polarized (a matter of about three standard deviations below the expected v/c value). However, three different methods [see Sec. III D (2), (3), and (5)] applied to these same betas have since yielded large (right-handed) helicity. Since we know there is nothing wrong with the ordinary Bhabha differential cross section (A54) which itself is highly sensitive to whether one takes into account the interference effects, it must be presumed that the Bhabha polarization method is intrinsically sound.

Transverse polarization analysis.—It is interesting to contrast the possibilities for transverse polarization analysis with the longitudinal case just discussed. Specific discussion is for the electron-electron case. One may examine again the basic Møller matrix element, and restrict again to 90° scattering angle in the center-of-mass system where the algebra is easiest and the physics most meaningful for our purposes. For the efficiency of a transversely polarized target electron we have the nonrelativistic value unity at sufficiently low energy. With increasing energy, the efficiency falls off very rapidly compared to the longitudinal case just described: if we average over azimuth of the plane of collision the efficiency ϵ is down to 0.5 at only 0.65-Mev laboratory energy, 0.3 at 1.25 Mev, 0.1 at about 3.4 Mev and ultimately becomes zero. If, instead, the plane of the collision is restricted to include the initial "spin vectors" involved (to include the spin vector of the polarized target electron will suffice) then the efficiency falls off almost as fast with energy, but does have a nonzero residue in the high-energy limit of $+0.11$. Or, if the plane of collision is required to be at right angles to this direction, ϵ saturates at -0.11 in the high-energy limit, having passed through zero at about 3-Mev laboratory energy.

For practical purposes, the transverse Møller method does not compare favorably with the longitudinal Møller method. The details for the transverse Bhabha method can be had by inspecting the matrix element again. A paper by Stehle (St58) lists in convenient form both Møller and Bhabha matrix elements.

Spin erection.—A further comment on Møller scattering relates to its hypothetical use as a spin erector (compare the electrostatic spin-precessor and the reprojection spin erector discussed in Sec. III B). The helicity memory through the Møller event can be rather large: given an incident longitudinal particle,

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

let it make a single collision in an unpolarized target. Then either of the offcoming particles, as viewed in the center-of-mass system will have a certain memory of the original helicity. Speaking as usual of 90° scattering, this efficiency for retention of helicity is +0.6 at 0.25 Mev, +0.8 at 0.6 Mev, and +0.9 at 2 Mev in the laboratory. The high-energy limit is 8/9. If then we transform either of the final electrons to the laboratory system we find that the over-all efficiency for transfer of longitudinal polarization P_p into transverse polarization P_s(lab) departs from zero $\sim 0.71E$, reaches the value 0.4 at 0.25 Mev and 0.5 at 1.0 Mev. It then falls with increasing laboratory energy E, approximately as $0.89(E+2\text{ Mev})^{-\frac{1}{2}}$, decreasing to 0.25 at 10 Mev. The direction of the resulting P_s is clear from the description—it points away from the line of collision for right-handed input. Longitudinal target polarization likewise could be transferred to P_s(lab) using the same numbers, given an unpolarized incident beam. Again, the transfer of polarization from P_p to P_p(lab) can be had from the transfer efficiencies just quoted by multiplying in each case by $(E+1\text{ Mev})^{\frac{1}{2}}$.

If one asks now for the transfer efficiency from P_s to P_p(lab) in a 90° Møller scattering act from an unpolarized target, which is the inverse of spin erection, one finds quite different numbers from the spin erection case. Thus in contrast to the spin-erectors already described: macroscopic electrostatic device, crossed fields device, or the use of atomic scattering for spin reprojection, this is an illustration of a nonreversible device. For $P_p \rightarrow P_s$(lab), the spin direction of the participating target electron and also of the conjugate recoil electron are (individually) about 90% certain. For $P_s \rightarrow P_p$(lab) these spin directions are about 90% uncertain, except at very low energy.

As for use of Bhabha scattering for spin erection or de-erection, the various transfer efficiencies at high energies will be nearly the same as for the Møller case. For low energies, even for 90° scattering some memory of "which one is the positron" will persist through the collision, and at very low energy the various transfer efficiencies at any angle will be apparent from the geometry of the collision. The discussion has by now become quite academic. For positrons, or indeed electrons, except at quite high energy one would generally be better off to use single, plural or multiple Coulomb scattering for de-erection of spin.

Finally, neither Møller nor Bhabha formulas show any right-left asymmetry as there is in Mott scattering, even allowing both target and beam to be (transversely) polarized.

(2) *Annihilation of Slow Positrons in a Ferromagnet*

Suppose that initially polarized positrons could be brought to rest (that is to essentially thermal energies) in a ferromagnetic material, and still retain some of the original polarization, so that finally $\langle\sigma_z\rangle \neq 0$. We consider complete polarization for simplicity. The ultimate fate of over 99% of these stopped positrons is two-quantum annihilation, and against electrons of a definite spin polarization, namely opposite to that of the positrons. The total annihilation gamma momentum represents, to the extent that the positrons themselves contribute negligible momentum, the momentum (Fourier transform of wave function) of those electrons susceptible annihilation (for example, La57).

In iron, one believes that the magnetic electrons (the "two" out of the 26 electrons per atom) surely must have different momentum components than the average electron. Thus, quite aside from positron polarization, there would have to be some correlation between amount of departure from collinearity of the two annihilation photons and whether or not a "magnetic electron" had been annihilated. Since the deeper bound electrons, which are not polarized, also are presumably not very accessible to a slow positron for annihilation, this may tend to enhance the correlation. What could be measured in the two-quantum angular correlation is the relative abundance of the higher momentum portion (thus, in the wings of the curve, say from 5 to 10×10^{-3} radians away from 180°) for the two analyzer settings: magnetization along $+z$ and along $-z$. Thus if $\langle\sigma_z\rangle$ for the positrons were positive, then + magnetization must yield more of the higher momentum events than − magnetization, *providing* the magnetic electrons tend to have higher momentum than the average electron with which a positron may annihilate.

This method was devised and used successfully by Hanna and Preston (Ha57). Their first polarization results showed (accepting that the magnetic electrons have generally higher momentum) that Cu^{64} positrons are right-handed. In later work (Hb57, Ha58) various ferromagnetic materials and a number of different positron emitters have been compared. Figure 10 shows the apparatus used in certain of the experiments of Hanna and Preston. Instead of using the more standard

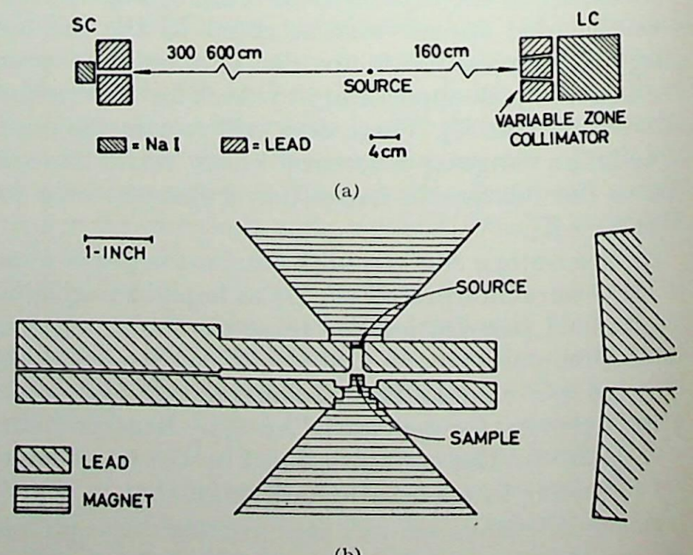


FIG. 10. Angular correlation apparatus of Hanna and Preston (Ha58) for measuring two-photon angular correlation from polarized positrons annihilating in a ferromagnet.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

momentum measurer as originally introduced by De-Benedetti and collaborators (D50) which essentially defines projected angle between photon momenta $-\mathbf{k}_1$ and $\mathbf{k}_2$, they used an annular detector slit located opposite to a point detector so that the actual angle θ between $-\mathbf{k}_1$ and $\mathbf{k}_2$ is defined. The particular choice of geometry for the gamma detection is not crucial for making polarization measurements however. If this technique eventually makes contact with any anisotropy of the momentum of the polarized electrons (with respect to crystal axes) then choice of detector geometry can be quite important.

The measured δ on reversing the state of magnetization, as obtained by Hanna and Preston, range from 5% to 10% depending on controllable circumstances, such as inserting absorbers between source and iron analyzer to "clean up" the positron beam.

There is no reason to suppose that the positrons on slowing have partaken of the polarization of the ambient electrons to any extent. To consider how annihilation-in-flight competes with the ordinary energy loss processes, the fact that a positron has survived so as to finally appear within tens of milliradians of 180° in the angular correlation predisposes it to be polarized, by a fraction of a percent, and parallel to the predominant electron spins. However, such added polarization would be the same for both states of magnetization (if referred to the magnetizing field direction) and therefore could not influence the measured δ to first order (and to any order, no matter how large such added polarization might be, it could not give a false sign to the measured δ).

As with analysis of positron polarization by means of positronium in the "plastics," absolute polarization measurement by the present technique awaits further development. But relative polarization measurement is feasible now.

The sign of the response of annihilation-at-rest in iron, the sign of the response of positronium in plastics, and the sign of the response of positronium in gases, are mutually consistent (the iron method has, for entirely different reasons, the same sign as for gases; the method using plastics has the opposite sign). Figure 11 shows a comparison plot between measured δ for iron (mild steel) and for a plastic (Teflon), run in the same apparatus at the same angular resolution. The resolution chosen (4×10^{-3} radian) seemed to be suited to the iron, but is known to be too small for optimum response in Teflon (III C). In any case the opposite response to the same positrons is evident. Hanna and Preston have checked (using Cu^{64} positrons) the response of polyethylene with respect to iron and it was opposite (Ha58).

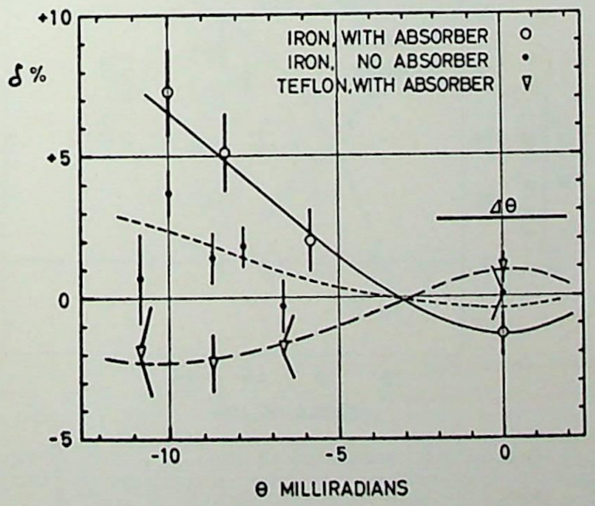


FIG. 11. Response of iron annihilator to polarized sodium-22 positrons compared to response of a plastic, both run at same resolution $\Delta\theta$ in same apparatus. Measured δ on reversing magnetic field is opposite in sign for the two materials. Positive δ means higher coincidence rate for H parallel to positron spins. (PO58).

The function of the absorber, which has been applied both at Argonne Laboratory and at Pittsburgh, is simply to enhance the measured δ, by selecting the higher energy portion of the actual positron spectrum which emerges from the source material plus the source backing. Thus, over and above the obvious fact that the effective v/c is slightly higher, clearly one is selecting particles of better polarization in the z direction. Depolarization of a positron beam in flight appears not to be a very serious matter.

A further comment on the three annihilation-at-rest methods for positron polarization measurement, is that it is generally quite difficult to join on in series a spin erector or de-erector, or an energy analyzer, principally because of the large ambient magnetic field in the immediate vicinity of the active region. For transverse polarization measurement, the annihilation-at-rest in a ferromagnet would clearly be the most easily adapted, since the magnetic field strength just outside the active sample need not be objectionably large.

(3) *Annihilation-in-Flight in Polarized Electron Target*

A method which is intrinsically nonrelativistic, but which can be used at fairly high energy nonetheless, is based primarily on the fact that at rather low energies, positron and electron helicities must be the same for two-quantum annihilation in flight to take place (three-quantum annihilation-in-flight is smaller by $\sim1/137$ and does not concern us here). There is, however, a strong departure from the nonrelativistic situation as regards efficiency for a longitudinally polarized analyzer to detect longitudinal beam polarization, at least assuming any and all two-quantum events to be recorded. This behavior is shown in Fig. 12, the solid curve. A crossover in sign of response occurs at about 4-Mev beam energy, so that at higher energies annihilation takes place more readily with spins parallel (positive efficiency ϵ).

However, as was realized by Frankel and co-workers (Fr57), the low-energy behavior (that is, $\epsilon=-1$) can be retained at any beam energy provided that the line-of-annihilation (in the center-of-mass system the two photons define a line of annihilation) lies exactly along the line of collision. In practice one chooses a sufficiently

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

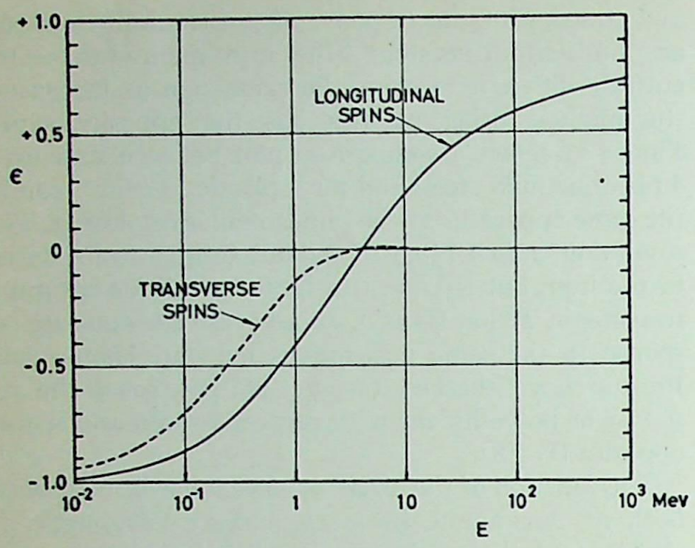


FIG. 12. Total cross section for annihilation-in-flight. Efficiency per polarized electron in target for analysis of longitudinal positron polarization (solid curve). Efficiency for transverse polarization analysis (dashed curve). Positive ϵ means parallel spins are favored to annihilate. From (P57).

narrow cone in the center-of-mass system that the high relativistic type events (those tending to make $\epsilon=+1$) are partially excluded from being counted. Thus, by a judicious choice between loss of efficiency for too large a cone and loss of counting rate for too narrow, a first-order experiment can be done on rather high-energy positrons. Frankel *et al.* used Ga^{66} (energies up to 4 Mev) with the result that these positrons are quite highly polarized, and right-handed, in line with the result of Deutsch *et al.* using another method [Sec. III D (5)]. In this type of experiment one necessarily runs into the oblique foil situation as in the Møller experiments (compare Fig. 9).

For energies much above 4 Mev, say 40 Mev, one would probably be better off to accept all annihilation events.

Transverse polarization.—The analog of the longitudinal case just discussed is mentioned briefly, for completeness sake, by examining the cross section for total annihilation-in-flight for a transversely magnetized foil. The efficiency per polarized electron is shown in Fig. 12 (as the dashed curve) as a function of positron energy in the laboratory. It is down to 0.5 by about 0.25 Mev and is essentially zero by about 3 Mev. Even restricting to the line of annihilation parallel to the line of collision, the efficiency falls off equally fast with increasing beam energy.

If the line of annihilation is restricted to be at right angles to the line of collision, the azimuthal dependence of the photon intensity becomes essentially $\cos^2\theta$ or $\sin^2\theta$ at high energy, depending on whether the input spins are parallel or antiparallel. This effect is about 50% efficient at 3 Mev and 90% efficient at 30 Mev (P57). Regarding counting rate such an approach would be expensive for two reasons: restricting the azimuth, and by restricting to cross-wise annihilation events, which unfortunately are fewer than the others (the ordinary Dirac differential annihilation cross section varies with beam energy so that a factor two is lost at 6 Mev, and a factor three at 20 Mev).

A second-order (hypothetical) method for transverse positron polarization analysis using annihilation-in-flight is discussed in context in Sec. III D (5).

(4) *Polarized Bremsstrahlung*

Longitudinal.—A spectacular and useful method for analyzing longitudinal polarization of energetic e^- and e^+ was devised largely intuitively by Goldhaber, Grodzins, and Sunyar. The method is to extract the longitudinal particle polarization information from the beam by means of the usual "forward" bremsstrahlung. This affords a natural, convenient, and (by polarization standards) copious source of circularly polarized photons. The secondary beam of photons is analyzed by comparison with atomic electrons (see Appendix C on the Compton method for photon helicity measurements).

Figure 13 shows the δ obtained by Goldhaber *et al.* (G57) on reversing the (transmission type) Compton analyzer through which the bremsstrahlung spectrum from Y^{90} betas was counted. These photons were found to be highly polarized. The Compton analyzer was here throwing in no spurious δ since the crossover in sign of the response is observed to come almost exactly where it should according to the Compton formulas, at 0.64 Mev.

Early calculations by McVoy (M57) showed that the transfer efficiency P_p to photon circular polarization P_k should be $\sim$0.95 at least for the truly forward photons, and near the tip of the bremsstrahlung spectrum. One of McVoy's curves (M57) implied**** that

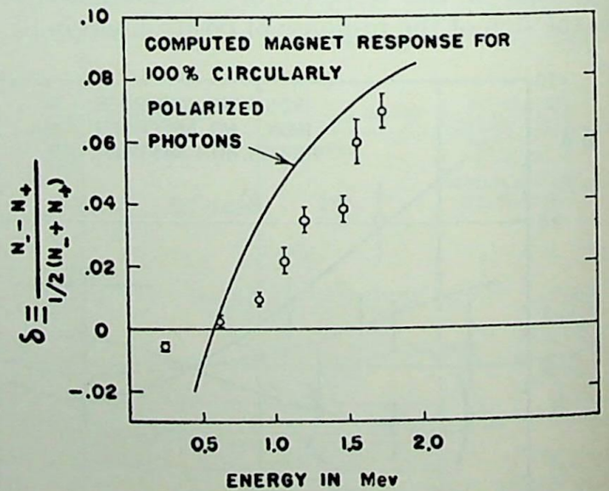


FIG. 13. Polarized bremsstrahlung from Y^{90} betas. Measured δ on reversing transmission type Compton analyzer as function of photon energy accepted. From Goldhaber *et al.* (G57).

**** See corrected curve (Mc58). The main effect in correcting the curve is that as $k/(E_1-1)$ decreases from unity the transfer efficiency begins to fall off considerably sooner than was at first believed, with the result that typically one should expect somewhat

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

such high efficiency should persist over a rather wide range of fractional photon energy, $k/(E_1-1)$. Later, bremsstrahlung computations were made over a range of photon angle as well as fractional photon energy by Fronsdal and Überall (F58). Computation on this problem has also been done by Claesson (C57) and by Böbel (B57). The problem of particle depolarization during such slowing as takes place before the bremsstrahlung act seems not to be an all-important problem, but must ultimately be faced.

On the basis of the first curve of McVoy (M57) some precise measurement of longitudinal particle polarization has been reported. For example, P^{32} betas were measured somewhat in the nature of a by-product in a series of experiments by Boehm and Wapstra (Bo58). Here the implied longitudinal polarization was -0.97 ± 0.06 of v/c. Until such time as (i) the bremsstrahlung method has been seriously checked, experiment *versus* theory, and (ii) the number of polarized electrons/cm^3 in a given ferromagnetic material under given conditions is truly known, polarization measurement by this method to absolute precision approaching 5% should be considered as somewhat on the optimistic side. As matters stand, item (i) cannot be divorced from item (ii), since the former is currently based upon the latter.

Aside from their intrinsic interest, bremsstrahlung sources are being widely used as calibrators for photon-

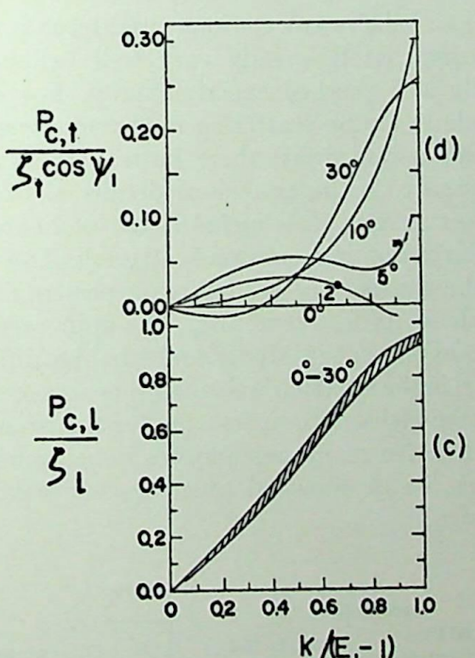


FIG. 14. Bremsstrahlung calculations. Theoretical transfer efficiency from longitudinal electron polarization to circular polarization of photon (lower curves) as function of fractional photon energy. Transfer from transverse particle polarization to photon helicity (upper curves). Photon emission angle θ is shown. Incident electron energy is 2.5 Mev. From Fronsdal and Überall (F58), with some details omitted.

less circular polarization, by perhaps 10 to 15%, unless the conditions correspond to the tips of both the beta spectrum and the bremsstrahlung spectrum (which is scarcely practical).

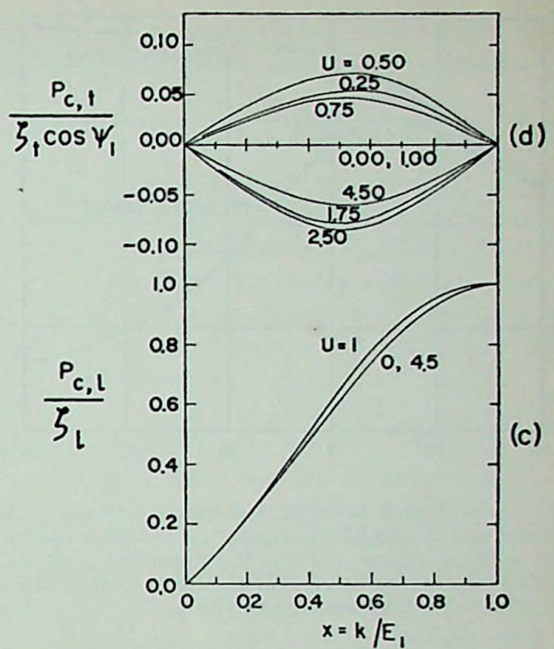


FIG. 15. Bremsstrahlung calculations. See also Fig. 14. Incident beam energy is 52 Mev. Photon emission angle in radians given by $U=103\theta$. From (F58).

helicity measuring devices. The bremsstrahlung process occupies a strategic position among the various polarized sources and measuring devices which eventually must come to be well understood.

Circular polarization of bremsstrahlung was the method used in the first successful helicity measurement of betas from muon decay. The experiment of Culligan *et al.* (Cu57) which yielded right-handed positrons, was the first suggestion that came from a *helicity* measurement that either lepton conservation or the tensor interaction in certain beta-decay processes would have to be given up. Since then, bremsstrahlung measurements on the helicity of both the negative and the positive betas from decay of the negative and positive muon, respectively, has been performed (Ma58). Both signs turn out to be the same as for the corresponding beta from nuclear beta decay, and the sign obtained by Culligan is thereby confirmed.††††

Typical results from the computation by Fronsdal and Überall (F58) are given in Figs. 14 and 15.‡‡‡‡ To a first approximation, all of the forward cone of bremsstrahlung has about the same transfer efficiency (from P_p to P_k) at a given fractional photon energy.

Transverse polarization.—The transfer efficiency from P_s to photon helicity P_k has been computed by Claesson (C57), by Böbel (B57), and by Fronsdal and Überall (F58). The efficiency here is considerably diminished from the longitudinal case, $P_p \rightarrow P_k$, as may be seen from the upper curves in both Figs. 14 and 15.

†††† See also Culligan, Frank, and Holt, Proc. Phys. Soc. (London) 73, 169 (1959) for further work.

‡‡‡‡ These authors are thanked for sending results prior to publication.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

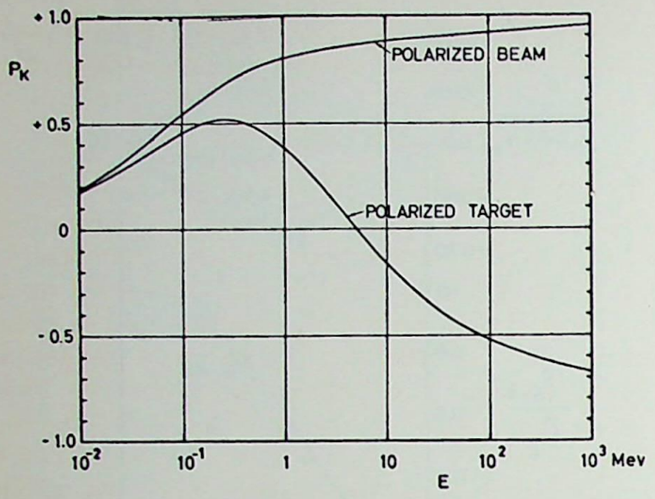


FIG. 16. Annihilation-in-flight. Computed transfer of helicity from positron to the annihilation quantum of higher energy in the laboratory system (upper curve). Right-handed positron yields right-handed photon (positive P_k). If target, instead of beam, is polarized, lower curve applies. From (P57).

(5) *Circular Polarization of Annihilation Photons*

As has been stressed before, there is no helicity information in the photons from nonrelativistic annihilation of positrons (nor is there in the photons which lie strictly along the line of collision, at any beam energy). However, for annihilation in flight at quite modest energies, the photon in the "opposite hemisphere" to whichever of the particles is longitudinally polarized takes on to some extent the same helicity as that particle had. This transfer efficiency is implied by the upper curve of Fig. 16, which specifically gives the circular polarization of the *higher* energy member of the pair of photons, per unit polarization of the beam. The target is assumed to be unpolarized. In actual circumstances, depending on geometry and energy discrimination in the photon detector, one may effectively accept only photons at center-of-mass angle $0<\theta<\theta_{\max}$, and usually with varying efficiency as a function of θ. What is shown here is simply the situation where all photons $0<\theta<90°$, and only those, are presumed to be analyzed (all counted and analyzed with equal efficiency). The process here is typically relativistic, and thus the higher the beam energy the more convenient the method becomes (with regard to energy discrimination, and geometrically). This may be compared with the experimental situation in Sec. III D (3), where a balance had to be maintained between competing effects.

The case where the target and not the beam is polarized is also indicated in Fig. 16, but is of no particular concern to us here. One sees however how the crossover in sign at 4 Mev relates to the similar crossover in the solid curve of Fig. 12.

Deutsch and collaborators (DS57) have used annihilation-in-flight in an unpolarized converter (Lucite) to measure the helicity of the positrons from Ga^{66} (the high-energy group has 4-Mev end point). Their apparatus is shown in Fig. 17. A selected portion of the beta spectrum is focused onto the converter, the high-energy (forward) photons being then analyzed for circular polarization in the transmission Compton analyzer shown. The computed δ, based on fully polarized (right-handed) positrons compared with the measured δ as a function of photon laboratory energy (for given input beta energy) implied essentially 100% longitudinal polarization. The measured values of δ ranged to 8%.

This second-order method has been used without preselecting the energy of the positrons by Boehm *et al.* (Bo57).

Transverse polarization.—For the process just considered (annihilation-in-flight against an unpolarized target) there is generally a nonzero transfer from any P_s in the particle beam to P_k of the annihilation radiation. This is a maximum at center-of-mass angle 90°, that is when the line of annihilation lies crosswise to the line of collision (P58). The magnitude of this efficiency generally exceeds the analogous transfer term for the bremsstrahlung process. At 3 Mev the transfer efficiency is a maximum of 0.55, and it is rather symmetric about 3 Mev on logarithmic energy scale, falling to half of this value at 0.3 Mev or at 60 Mev. This hypothetical method suffers from intensity limitations for the same two reasons mentioned in Sec. III D (3).

IV. CONCLUDING REMARKS

A. On Calculations

So far, all helicity calculations cited have proceeded from theory that is already very well believed, and reasonably well checked experimentally. For example (to include Compton scattering for completeness), the "spin effects" were always there, both theoretically and experimentally, in the nonhelical details of the differential cross section. This includes loss of definition of plane polarization in Compton scattering. The same is true of the existence of ground-state positronium and the details of its fine structure. The spin terms were known to be present in Møller and Bhabha differential scattering, in the annihilation-in-flight cross section, and in the Bethe-Heitler approximation for the bremsstrahlung process. No really new physics is being used here in making use of standard processes for polarization measurement.

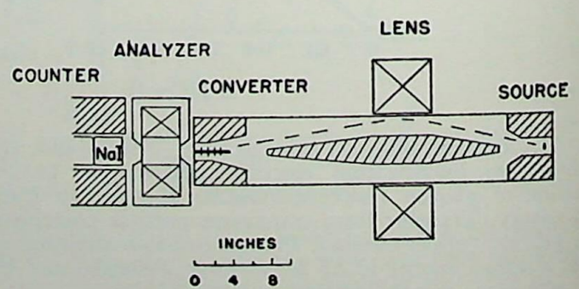


FIG. 17. Experiment to measure positron helicity by transferring polarization information to annihilation (in flight) photon, which is analyzed for helicity in Compton analyzer. From Deutsch and colloborators (DS57).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

In order to make practical use of these spin effects, lists of matrix elements (transition amplitudes) from fully specified state to fully specified state are good things to have accessible. Then none of the polarization implications in the original theory are averaged out until the people devising a particular experiment are ready to do so.

For fast particles, transverse polarization is generally less amenable to measurement than is longitudinal polarization. Perhaps it is not too early to begin looking beyond those polarization effects which are already in the extant theory. Under this heading would come an enquiry into the possibilities of right-left asymmetry in bremsstrahlung intensity from transversely polarized particles. Here one would have to go beyond the Bethe-Heitler approximation, reminiscent of how, in Coulomb scattering without radiation, one has to go beyond the first Born approximation in order to find any polarization effects other than "reprojection" of spin.

B. On Experiment

Of great assistance in succeeding with the experiments discussed here has been the fact that all the methods are really *null* methods. In these experiments, a large number of parameters such as solid angle, energy straggle, backgrounds, are arranged to be *the same* for both "states" of the analyzer. By and large it is nowadays as easy to demonstrate spin-helicity effects, using the large input polarizations now known to be available from beta decay, as it is to demonstrate in a nonpolarization experiment some relativistic feature of a differential cross section, or to measure an absolute cross section, to compare with Dirac theory. The reason is that none of these latter experiments can be made into a null method. In exploring details of a differential cross section there is always some important parameter that changes: the energy of a photon, an angle of obliquity in a scattering foil, signal-to-noise ratio. These things that matter will change always in a systematic way with the independent variable of the particular experiment. Hopefully, these systematic changes are computable, measurable or otherwise understood. As for absolute cross-section determinations (except perhaps for total Compton cross section), these are notorious for having many factors which must be controlled.

With the labor and care necessary for say a 10% measurement of differential cross section, one now gets typically a 1% measurement on a change in counting rate which relates linearly to the polarization being measured. The intrinsic reliablity and power of the balancing or null approach makes this possible. Suppose the polarization effect causes a 10% change in counting rate, then one would have a measurement of polarization good to 10%. For smaller output asymmetries, one must try to balance more thoroughly. In any case, it appears that the potential power of the null methods needs to be stressed, especially in telling of polarization measurement to persons outside the immediate field.

C. On Calibration

We look forward to the time when a number of standard (that is, reproducible) beta and bremsstrahlung sources and analyzers are available. Numbers will have to be attached, these numbers will have to be cross-checked with each other somehow, and of course "theory" and the numbers will have to be compatible.

Aside from precision "numerical" calibration which is the course presently being followed, there is a different approach which might be mentioned. Just as an unpolarized beam has no polarized part, a fully polarized beam has no unpolarized part. This is not trivial: It should be recalled that if the problem were to demonstrate that a certain beam is unpolarized to within an infinitesimal, perhaps 0.5%, then a polarization analyzer whose efficiency is known only crudely (say within a factor two) might suffice—one performs a null experiment. In dealing with degrees of polarization P which are much closer to unity than to zero, it might be worthwhile to consider possibilities for measuring $1-|P|$, instead of P itself. Thus, if $|P|$ were ~ 0.97, a measurement on $1-|P|$ good to only 30% is equivalent to a measurement on $|P|$ good to one percent. The sign of P would of course be known before one embarked on this hypothetical precision experiment. Admittedly, such experiments would have to be of second order and therefore quite difficult. Yet, it would be aesthetically pleasing to have proof of, say, 100% polarization presented in the form of a null experiment, to complement the more traditional approach which would hinge upon a chain of calibration numbers.

V. ACKNOWLEDGMENTS

The major part of this paper was written at the Institute of Physics, Uppsala. The author thanks Professor Kai Siegbahn for the hospitality of the Institute. A number of general discussions with Professor Mael Melvin at Uppsala have been appreciated. Association in past years with Professor Philip Stehle of Pittsburgh has laid some basis toward understanding of polarization problems.

The John Simon Guggenheim Memorial Foundation is thanked for a fellowship, and the further support provided by the Office of Ordnance Research, U. S. Army, is acknowledged.

APPENDIX A. BETA DECAY AND POLARIZATION

Allowed Fermi Decay

The connection between neutrino polarization and beta polarization can be illustrated in simple sketches (Fig. 18). We assume a pure V (vector) interaction, and let a positron be emitted for illustration. The other

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

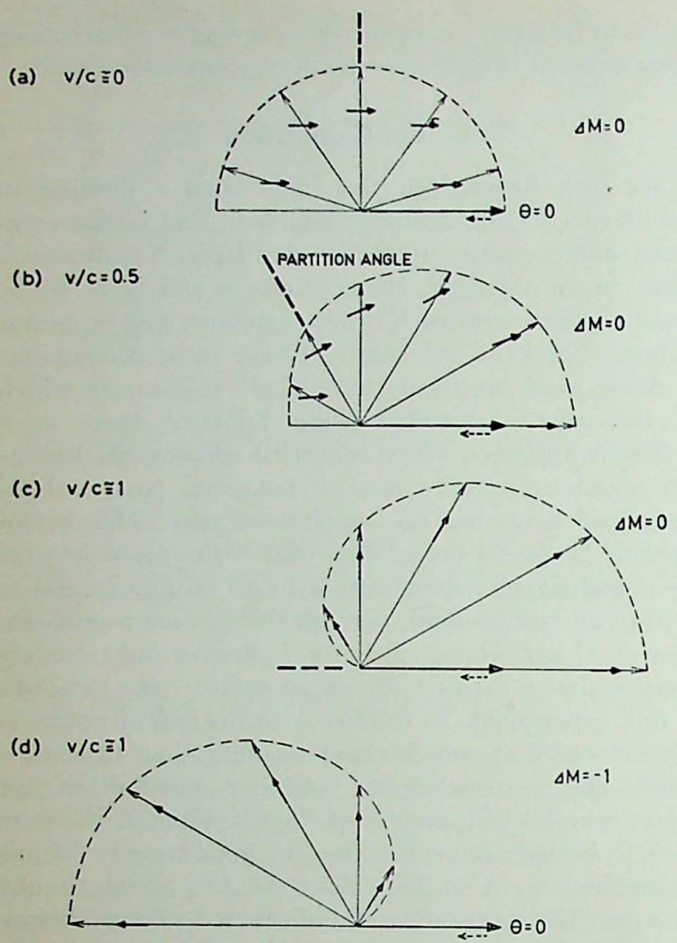


FIG. 18. Relations between polarization vectors and momentum vectors. The emitted neutrino momentum **q** defines direction $\theta=0$. Length of a given radius vector to the dotted envelope is proportional to the probability that the beta momentum **p** shall be in that direction. Spin-polarization vectors are shown by short arrows. Beta-polarization arrows (each representing degree of polarization unity) generally have meaning only in the rest system of the beta particle. Neutrino helicity -1 is used for illustration. Sketches apply for pure V or pure A interaction.

particle emitted is called a neutrino and not an anti-neutrino. In sketches (a), (b), and (c) the neutrino proceeds along the line $\theta=0$, to our right. Suppose it is left-handed. The sketches show (i) the angular correlation of the beta momentum **p**, with respect to neutrino momentum **q** (the envelope drawn is the relative probability of beta emission for given angle of emission, to be multiplied by d^3p); (ii) the polarization vector in the *rest system* of the beta particle, where the magnitude of the polarization is always unity since in this situation no degree of freedom remains.

These sketches rest on the fact that the *amplitude* for the beta to be at angle θ and to be right-handed is $(1+b)\cos(\theta/2)$, while the (coherent amplitude for it to be left-handed is $(1-b)\sin(\theta/2)$, where $b=p/(\gamma+1)$ and γ means total energy of the beta. It follows that the beta-neutrino correlation (unpolarized) is

$$W(\theta)=1+\frac{2b}{1+b^2}\cos\theta=1+\frac{v}{c}\cos\theta. \qquad \text{(A1)}$$

Alternatively from these amplitudes, the longitudinal polarization is

$$P_p(\theta)=\frac{(v/c)+\cos\theta}{1+(v/c)\cos\theta}. \qquad \text{(A2)}$$

Beginning again with these amplitudes, the longitudinal polarization, *independent of* θ is $+v/c$. We could have reversed the roles of beta and neutrino by requiring say a right-handed beta, then the angular correlation of the neutrino with this beta is just $(1+\cos\theta)$. A neutrino of zero rest mass is taken here.

Expression (A2) shows there is a "partition angle" at which the beta polarization is exactly transverse. This angle, defined by $(v/c)+\cos\theta=0$, moves closer to 180° as v/c increases. The drawings show how P_p tends to unity in the high-energy limit.

To pass from pure V interaction to pure S (scalar) interaction, one may reverse **q** and the neutrino helicity and otherwise keep the same patterns as sketched in (a), (b), and (c) of Fig. 18.

Nowhere in this discussion was anything said about nonconservation of parity: We could argue that what was plotted or given in the foregoing expressions was intended to represent only *half* of the Einstein-Podolsky situation. Then, by supplying the recipe for the *other* helicity of the neutrino, which says to reverse all polarization vectors, the picture would be completed. [Expression (A1) would be unchanged.]

It seems that only one-half of the picture occurs in nature, but this need have nothing to do with our trying to illustrate the connection between the momenta and helicities. It does supply the motive and explains our choice of *longitudinally* polarized neutrino in these figures.

Allowed Gamow-Teller Decay

In the foregoing section, although we pretended to be discussing only the pure Fermi case, drawings (a), (b), and (c) apply equally well to pure GT—to the $\Delta M=0$ part at least. For the A (axial-vector) interaction we may keep the drawings as they stand, for pure T (tensor) interaction we can reverse **q** and the neutrino helicity. To depict the situation for an offgoing right-handed neutrino, reverse all polarization vectors. To complete the picture for GT, a $\Delta M=-1$ transition is sketched as part (d) of Fig. 18, only the high-energy limit being shown for illustration. A $\Delta M=+1$ transition would be represented by reflecting picture (d) in the plane $\theta=90°$ (including the polarization arrows). Throughout we have been taking the axis of quantization for J to be along $\theta=0$.

As regards the unpolarized angular correlation, we now have competing parts: the angular correlation for a picture like (b) is $1+v/c\cos\theta$, and for a picture like (d) is $1-v/c\cos\theta$. However, transition (d) occurs twice as often as (b), at least for unpolarized initial and final nucleus. By not forcing a definite helicity on the neutrino, $|\Delta M|=1$ transitions occur twice as often as

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

$\Delta M=0$ transitions, as required by spectroscopic stability. Or, requiring definite helicity for the neutrino, picture (b) represents "half" of a triplet $M=0$ state (of neutrino plus beta). Thus, either way we look at it, the angular correlation is the same,

$$W(\theta)=1-\frac{v}{c}\cos\theta. \tag{A3}$$

Picture (a) and a nonrelativistic version of (d) are what apply in a pure GT K-capture event. With respect to $\mathbf{q}$, the ΔM imparted to the nucleus is either -1 or zero (for a left-handed neutrino emitted), the former transition occuring with twice the probability of the latter.

Now if we require, for the first time, a given "ΔM" in the laboratory, the neutrino will seldom proceed along the axis of quantization, and simple sketches such as these are no longer feasible. So many important experiments are related to J_z in the laboratory that the discussion should be pursued a little further. Inspection of equation (1) of Jackson *et al.* (J57) shows, for pure A interaction, say, that the $\Delta M=0$ transition has angular correlation (z axis fixed in the laboratory)

$$1-\frac{\mathbf{p}\cdot\mathbf{q}}{E_eE_\nu}+2\frac{p_zq_z}{E_eE_\nu}. \tag{A4}$$

Of course, $\mathbf{p}$ and $\mathbf{q}$ *separately* are isotropic in the laboratory since one cannot "align" a spin-$\frac{1}{2}$ particle but can only polarize it. From (A4) the correlation between $\mathbf{p}$ and $\mathbf{q}$ can be readily worked out to be $1-\mathbf{p}\cdot\mathbf{q}/3E_eE_\nu$, which is the same as (A3).

For comparison the $\Delta M=-1$ correlation [again from (J57)] is

$$\left(1-\frac{p_z}{E_e}\right)\left(1+\frac{q_z}{E_\nu}\right) \tag{A5}$$

and, although $\mathbf{p}$ and $\mathbf{q}$ are now each tied closely to the z axis, the correlation between them is the same as for the $\Delta M=0$ case just treated.

If we had the simple case J, $M=1$, -1, making a transition to $J=0$, the neutrino (taken to be left-handed) would be emitted as $1+\cos\theta$, the beta (independently) as $1-v/c\cos\theta$, the factor $\frac{1}{3}$ being now absent since only one possibility for ΔM has been allowed. On the other hand, if we ask for the polarization of daughter nucleus (let us now go from $J=0$ to $J=1$ for illustration) having found a beta along z, then the $\Delta M=0$ contribution is again in competition with $|\Delta M|=1$, and the factor $\frac{1}{3}$ often associated with GT transitions reappears. The beta—circularly polarized gamma correlation is treated in detail by Alder *et al.* (A57) for example.

APPENDIX B. SHORT COMPENDIUM OF HELICITY SIGNS

The experiment of Wu *et al.* (Wu57) showed that the negative beta particles from polarized Co^{60} nuclei come off predominantly opposed to the field H. Thus these betas as predominantly opposed to the direction of the original nuclear spin, because ground-state Co^{60} has a positive g value (W55). This decay has a decreasing spin change ($5\rightarrow 4$) and therefore the beta and the emitted antineutrino jointly carry off angular momentum parallel to the original nuclear spin, unless we suppose that conservation of angular momentum is in question. Since none of this angular momentum carried off is orbital (this is an allowed transition) it has to be carried entirely by the spins of the two particles. Suppose, for argument, that the original nuclear polarization were complete. Then the two particles must carry off one unit of angular momentum referred to the z axis, the field axis. They are both spin-$\frac{1}{2}$ particles and so must "aid each other." The Wu result implies then that the betas from Co^{60} have to be predominantly left-handed. To summarize, granting the sign of g, the sense of the spin sequence here, and granting the conservation of angular momentum, the "sign" that emerges from the Wu experiment is just the helicity of those betas.

Soon after the Wu experiment, the left-handed helicity of the Co^{60} betas was checked by Frauenfelder and collaborators (Fa57) by means of Mott scattering. Page and Heinberg found that positrons from Na^{22} were predominantly right-handed (PH57). Simultaneously Schopper (S57) showed that Na^{22} and Co^{60} (both having the same sense in the spin sequence) emitted photons of opposite handedness, when measured at nominal 180° to the line of the emitted beta, the sign for Co^{60} being in accord with the Wu experiment. Ga^{66} believed to decay via a pure Fermi transition was shown by Deutsch and co-workers (DS57) to give highly polarized right-handed positrons. [No attempt is made to record many of the early polarization experiments which quickly extended the list to include, for example, mixed (F and GT) transitions, forbidden transitions, etc.]

Strictly, none of these experiments shows anything about neutrino helicity (that is, whether it is left or right), since the neutrino direction is not measured. What is measured in the three kinds of experiment just mentioned, Wu experiment, polarized beta experiment, and beta—circularly polarized gamma experiment, is a a sense of circulation and coincidentally a *beta* direction, nothing more.

The helicity of the neutrino emitted in K capture (presumably the same particle as the one emitted together with a positive beta) was measured *directly* by Goldhaber *et al.* (G58) and found to be left-handed.§§§§ One concedes again conservation of angular momentum and certain j values of the levels involved.

§§§§ The sign and order of magnitude of the neutrino helicity for the decay used by Goldhaber ($Eu^{152m}\rightarrow Sm^{152*}$) has now been checked at Uppsala (M58).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

To review further would lead into the status of beta-decay theory. The algebraic signs relating to a small and specific set of decays are given, signs which have not been in jeopardy and presumably will stand. This is done for the benefit of any person who may have had the notion that all had been confusion with regard to helicity measurement.

APPENDIX C. COMPTON ANALYSIS OF PHOTON HELICITY

The use of a Compton analyzer for circularly polarized radiation rests on a spin-helicity correlation present in the matrix element of Klein and Nishina. Details‖‖‖‖‖‖ of the spin-helicity part of differential or total cross section are given in (G53), or see the review article (T). The Compton analyzers may be divided into two types as indicated in the following.

Transmission Type

Here the spin-polarization sensitivity of the *total* Compton cross section is utilized. The analyzing efficiency, per polarized "iron" electron, has a faint unexplored maximum of 0.06 at about 0.2 Mev, and is zero at 0.64 Mev. From this energy to 14 Mev, where the efficiency is 0.4, it increases almost as the logarithm of $E/0.64$ Mev, and at very high energy becomes unity. Before helicity measurement became of general interest, this type was used in second order (that is, the two sets of spins in an iron analyzer acted as the polarizer-analyzer combination to verify that the spin-helicity term could be seen experimentally). A transmission analyzer was first used in first order, to measure degree (and sense) of polarization in a beam, by Trumpy (T57). This was a "nonparity" experiment, as it used capture gamma rays produced by prepolarized slow neutrons. The first application of this analyzer in a "parity" experiment was by Goldhaber *et al.* (G57), to demonstrate and measure the polarization of bremsstrahlung by Y^{90} beta rays. Lundby and co-workers have used it in beta—circularly polarized gamma correlation work (L57).

Owing to the inescapable null in the efficiency curve at 0.64 Mev, the transmission analyzer is best suited for rather high energy work, but has been used successfully on a gamma line as low as 0.9 Mev (G58).¶¶¶¶ Normally, a gamma ray is not seriously degraded in energy in passing through the iron absorber. Finally, this type of analyzer has zero sensitivity to any competing *plane* polarization in the photon beam.

Differential Type

In this type one counts, at a certain angle, photons scattered by polarized electrons. Some years before parity work, the differential analyzer was proven at Leyden (W55) on the Ni^{60} gamma rays following the beta decay of polarized Co^{60}. The method has since been revived by Schopper (S57) and by Boehm and Wapstra (BW57) for use in beta—circularly polarized gamma correlation measurements. Scattering at very large angles has recently been employed by Bernardini *et al.* (Be58) to measure the helicity of inner-bremsstrahlung. Unless one is prepared to make absolute scattering measurements (which are quite out of the question, compare Sec. IV B) a differential analyzer is quite sensitive to *plane* polarization,***** this sensitivity being independent of spin direction. Therefore, for a given geometry of source, scatterer and photon detector, a flat statement of "circular polarization efficiency" in terms of *fractional* change in intensity of scattered photons on reversing magnetization cannot strictly be given, unless some proviso is made regarding ellipticity (or planeness). This feature was at once recognized by the Leyden group, but appears to have largely escaped mention since.

Owing to counting rate considerations the scattering geometry is often such that any lines in the incident gamma spectrum are quite spread in energy as seen by the detector.

Bibliography

(H) W. Heitler, *Quantum Theory of Radiation* (Oxford University Press, 1944), second edition, or later.
(JR) J. M. Jauch and F. Rohrlich, *Theory of Photons and Electrons* (Addison-Wesley Press, Cambridge, 1955).
(KS) *Beta- and Gamma-ray Spectroscopy*, Kai Siegbahn, editor, (North-Holland Publishing Company, Amsterdam, 1955).
(MM) N. F. Mott and H. S. W. Massey, *The Theory of Atomic Collisions* (Oxford University Press, New York, 1949).
(S) L. I. Schiff, *Quantum Mechanics* (McGraw-Hill Book Company, New York, 1949), or later editions.
(SBD) Schweber, Bethe, and DeHoffmann (Row, Peterson and Company, Evanston, Illinois, 1955), Vol. I.
(T) H. A. Tolhoek, Revs. Modern Phys. **28**, 277 (1956).

A54 Ashkin, Page, and Woodward, Phys. Rev. **94**, 357 (1954).
A57 Alder, Stech, and Winther, Phys. Rev. **107**, 728 (1957).
A58 Alikhanov, Eliseiev, Lubimov, and Ershler, Nuclear Phys. **5**, 588 (1958).
B38 W. Bleakney and J. A. Hipple, Phys. Rev. **53**, 521 (1938).
B53 R. E. Bell and R. L. Graham, Phys. Rev. **90**, 644 (1953).
B57 G. Böbel, Nuovo cimento **6**, 1241 (1957).
Bi57 A. Bincer, Phys. Rev. **107**, 1431 (1957).
Bo57 Boehm, Novey, Barnes, and Stech, Phys. Rev. **108**, 1497 (1957).
BW57 F. Boehm and A. H. Wapstra, Phys. Rev. **106**, 1364 (1957).
B58 Bernardini, Brovetto, and Ferroni, Nuclear Phys. **8**, 294 (1958).
Be58 Bernardini, Brovetto, DeBenedetti, and Ferroni, Nuovo cimento **7**, 416 (1958).
Bo58 F. Boehm and A. H. Wapstra, Phys. Rev. **109**, 456 (1958).
BW58 Benczer-Koller, Schwarzschild, Vise, and Wu, Phys. Rev. **109**, 85 (1958).
C57 A. Claesson, Arkiv Fysik **12**, 569 (1957).
Cu57 Culligan, Frank, Holt, Kluyver, and Massam, Nature **180**, 751 (1957).

‖‖‖‖‖‖ For reasons which are not clear, expressions applicable for experimental design are occasionally attributed to an older publication (F38).

¶¶¶¶ Also compare Fig. 13.

***** In a beta-gamma cascade exhibiting nontrivial angular correlation, or where for any reason the gammas to be analyzed may have nonzero plane polarization referred to the plane of the Compton scattering act, this effect could conceivably affect numerical reduction of the data.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

CR57 Cavanagh, Turner, Coleman, Gard, and Ridley, Phil. Mag. **2**, 1105 (1957).
C58 M. Carrassi, Nuovo cimento **7**, 524 (1958).
D50 DeBenedetti, Cowan, Konneker, and Primakoff, Phys. Rev. **77**, 205 (1950).
D52 S. DeBenedetti and H. J. Richings, Phys. Rev. **85**, 377 (1952).
D53 M. Deutsch, *Progress in Nuclear Physics* (Butterworths-Springer, London, 1953), Vol. 3, p. 131.
D54 S. DeBenedetti and H. C. Corben, Ann. Revs. Nuclear Sci. **4**, 191 (1954).
D56 M. Deutsch, International Conference on Quantum Interactions of the Free Electron, University of Maryland (1956).
De57 DeShalit, Kupermann, Lipkin, and Rotham, Phys. Rev. **107**, 1459 (1957).
DP57 H. deWaard and O. J. Poppema, Physica **23**, 597 (1957).
DS57 Deutsch, Gittleman, Bauer, Grodzins, and Sunyar, Phys. Rev. **107**, 1733 (1957).
D58 H. G. Dehmelt, Phys. Rev. **109**, 381 (1958).
F38 W. Franz, Ann. Physik **33**, 689 (1938).
F57 U. Fano, Revs. Modern Phys. **29**, 74 (1957).
Fa57 Frauenfelder, Bobone, VonGoeler, Levine, Lewis, Peacock, Rossi, and DePasquali, Phys. Rev. **106**, 386 (1957).
Fb57 Frauenfelder, Hanson, Levine, Rossi, and DePasquali, Phys. Rev. **107**, 643 (1957).
Fc57 Frauenfelder, Bobone, VonGoeler, Levine, Lewis, Peacock, Rossi, and DePasquali, Phys. Rev. **107**, 909 (1957).
Fd57 Frauenfelder, Hanson, Levine, Rossi, and DePasquali, Phys. Rev. **107**, 910 (1957).
Fr57 Frankel, Hansen, Nathan, and Temmer, Phys. Rev. **108**, 1099 (1957).
F58 C. Fronsdal and H. Überall, Phys. Rev. **111**, 580 (1958).
G53 S. B. Gunst and L. A. Page, Phys. Rev. **92**, 970 (1953).
G54 R. L. Graham and A. T. Stewart, Can. J. Phys. **32**, 678 (1954).
G57 Goldhaber, Grodzins, and Sunyar, Phys. Rev. **106**, 826 (1957).
GB57 R. E. Green and R. E. Bell, Can. J. Phys. **35**, 398 (1957).
G58 Goldhaber, Grodzins, and Sunyar, Phys. Rev. **109**, 1015 (1958).
H56 M. Heinberg and L. A. Page, Bull. Am. Phys. Soc. Ser. II, **I**, 168 (1956).
Ha57 S. S. Hanna and R. S. Preston, Phys. Rev. **106**, 1363 (1957).
Hb57 S. S. Hanna and R. S. Preston, Phys. Rev. **108**, 160 (1957).
He57 M. Heinberg and L. A. Page, Phys. Rev. **107**, 1589 (1957).
Ha58 S. S. Hanna and R. S. Preston, Phys. Rev. **109**, 716 (1958).
J57 Jackson, Treiman, and Wyld, Phys. Rev. **106**, 517 (1957).
L57 Lundby, Patro, and Stroot, Nuovo cimento **6**, 745 (1957).
La57 G. Lang and S. DeBenedetti, Phys. Rev. **108**, 914 (1957).
M29 N. F. Mott, Proc. Roy. Soc. (London) **A124**, 425 (1929).
M32 C. Møller, Ann. Physik **14**, 568 (1932).
M55 H. Mendlowitz and K. M. Case, Phys. Rev. **97**, 33 (1955).
M57 K. McVoy, Phys. Rev. **106**, 828 (1957).
Ma58 Macq, Crowe, and Haddock, Conference on Weak Interactions, Gatlinburg (1958).
M58 I. Marklund and L. A. Page, Nuclear Phys. **9**, 88 (1958).
Mc58 K. McVoy, Phys. Rev. **110**, 1484 (1958).
P51 L. A. Page, Phys. Rev. **81**, 1062 (1951).
P54 T. A. Pond, Phys. Rev. **93**, 478 (1954).
P55 Page, Heinberg, Wallace, and Trout, Phys. Rev. **98**, 206 (1955).
Pa55 Page, Heinberg, Wallace, and Trout, Phys. Rev. **99**, 665(A) (1955).
PH56 L. A. Page and M. Heinberg, Phys. Rev. **102**, 1545 (1956).
P57 L. A. Page, Phys. Rev. **106**, 394 (1957).
PH57 L. A. Page and M. Heinberg, Phys. Rev. **106**, 1220 (1957).
Po57 T. A. Pond, postdeadline paper, Washington American Physical Society Meeting (1957).
PO57 L. A. Page and F. E. Obenshain, Bull. Am. Phys. Soc. Ser. II, **2**, 260 (1957).
P58 L. A. Page, Phys. Rev. **109**, 2215 (1958).
PO58 L. A. Page and F. E. Obenshain, final O. O. R. report, University of Pittsburgh (November, 1958) (unpublished).
S55 A. T. Stewart, Phys. Rev. **99**, 594 (1955).
S56 N. Sherman, Phys. Rev. **103**, 1601 (1956).
S57 H. Schopper, Phil. Mag. **2**, 710 (1957).
S58 H. Schopper, Nuclear Instr. **3**, 158 (1958).
St58 P. Stehle, Phys. Rev. **110**, 1458 (1958).
T51 H. A. Tolhoek and S. R. DeGroot, Physica **17**, 1 (1951).
T56 Telegdi, Sens, Yovanovitch, and Warshaw, Phys. Rev. **104**, 867 (1956).
T57 G. Trumpy, Nuclear Phys. **2**, 664 (1957).
Ta57 L. J. Tassie, Phys. Rev. **107**, 1452 (1957).
W55 Wheatley, Huiskamp, Diddens, Steenland, and Tolhoek, Physica **21**, 841 (1955).
W57 S. D. Warshaw, Phys. Rev. **108**, 713 (1957).
Wu57 Wu, Ambler, Hayward, Hoppes, and Hudson, Phys. Rev. **105**, 1413 (1957).
W58 C. S. Wu, *Proceedings of the Rehovoth Conference on Nuclear Structure*, H. J. Lipkin, editor (North-Holland Publishing Company, Amsterdam, 1958), p. 380.
Y50 C. N. Yang, Phys. Rev. **77**, 242 (1950).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

Invited Papers from the Conference on Weak Interactions Held at Gatlinburg, Tennessee, October 27-29, 1958

Introductory Note

SOMEWHAT more than two years have passed since the realization came to physicists that the reflection law known as the conservation of parity possesses limits in the generality of its application. A paradox in the decay of the τ mesons had led to some daring theoretical speculation, promoting direct laboratory tests which led in crucial experiments of rare, illuminating character to a re-direction of thought concerning those interactions which we call "weak." For these physicists, the assumption of conservation of parity suddenly lost its sacredness. The result was a flurry of action; experiments of new types were started, and theoretical channels were explored along new lines. Last summer, new observations were being reported in abundance, and it occurred to members of the Physics Division of the Oak Ridge National Laboratory that a meeting of those involved and interested in the subject might be of profit. A local conference committee was established,* the American Physical Society was approached and granted its endorsement and help in the welcome form of two additional members for the Committee,† and the conference took place at Gatlinburg, Tennessee, on October 27, 28, and 29, 1958.

The conference comprised twelve sessions, in which the guidance of the following experienced contributors to the field, acting as chairmen, was a vitalizing influence: M. Goldhaber, M. Deutsch, M. E. Rose, J. M. Robson, E. J. Konopinski, E. C. G. Sudarshan, S. B. Treiman, H. L. Anderson, R. W. Thompson, R. G. Sachs, and E. P. Wigner. Abstracts of the contributed papers have been published in the *Bulletin of the American Physical Society* (January, 1959) and their position in the total conference can be gauged by their numbers in relation to the summary of sessions given below. The substance of the invited papers is set forth in the pages that follow.

ARTHUR H. SNELL
Chairman

Summary of the Conference Sessions

Session A.	Parity Experiments. *Invited paper by* C. S. WU.
Session B.	Polarization following Beta Decay. *Contributed papers.*
Session C.	Experimental Beta Decay. *Invited paper by* J. S. ALLEN; *contributed papers.*
Session D.	Beta Decay of the Free Neutron; Time Reversal. *Invited paper by* V. L. TELEGDI; *contributed papers.*
Session E.	Theoretical Beta Decay. *Invited paper by* R. P. FEYNMAN; *contributed papers.*
Session F.	Theoretical π and μ Decay and μ Capture. *Invited papers by* M. L. GOLDBERGER *and* H. PRIMAKOFF.
Session G.	π and μ Decay and μ Capture. *Contributed papers.*
Session GA.	Recent Observations of π-e Decay (unscheduled session).
Session H.	μ Mesons—Experimental. *Invited paper by* R. L. GARWIN; *contributed papers.*
Session J.	Strange Particles—Experimental. *Invited paper by* J. STEINBERGER; *contributed papers.*
Session K.	Strange Particles and Weak Interaction—Theoretical. *Invited paper by* R. H. DALITZ; *contributed papers.*
Session L.	Weak Interaction—Summary. *Invited paper by* M. GELL-MANN.

* Members: J. L. Fowler, A. I. Galonsky, M. E. Rose, A. H. Snell, T. A. Welton, and H. B. Willard.
† C. S. Wu and E. J. Konopinski.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

REVIEWS OF MODERN PHYSICS VOLUME 31, NUMBER 3 JULY, 1959

Parity Experiments in Beta Decays*

C. S. Wu

Columbia University, New York, New York

INTRODUCTION

THE frontier of parity study has now advanced to the field of strange particles. The atmosphere in the field of beta decay appears unusually calm and quiet after the storm. I will try to piece together the jigsaw picture and to see what sorts of puzzles in beta decay have fallen into shape. Most urgent of all is the question, whether there are still any missing pieces, and if there are, what are they?

In principle, parity experiments are simply experiments designed to study the screw senses of particles, decay processes, or nuclear reactions. This all originated from our concept of left and right symmetry; that is, left and right are indiscernible.[1] To scientific minds, from the time of Leibniz to 1957, there was no inner difference, no polarity between left and right. The inner structure of space does not permit us, except by arbitrary choice, to distinguish a left from a right screw. However, a right screw or a right spinning particle shows up as a left screw or a left spinning particle in a mirror reflection. So from the right-left symmetry we arrived at the space-reflection symmetry. This symmetry implies that if a particle exists, the one obtained by reflecting it in a mirror must also exist. If a decay process can take place, the one seen in a mirror is also a possible one. The symmetry of space-reflection clearly requires that events or particles do not exhibit definite screw senses. That is why all parity experiments are concerned with the measuring of the screw senses. This screw sense is now christened with the elegant name of helicity and is defined as $\boldsymbol{\sigma}\cdot\mathbf{p}/|p|$, the spin along p.

The first parity experiment[2] using polarized Co^{60} was designed to test whether there are any screw senses in beta decay. The essence of the experiment was to line up the spins of the Co^{60} nuclei along the same axis and then to determine whether the beta particles were emitted preferentially in one direction or the other along the axis. The results showed that the electrons were emitted preferentially in the direction opposite to that of nuclear spin and therefore conclusively proved that the beta decay of Co^{60} behaves like a lefthanded screw or possesses a negative helicity. So parity is not conserved in beta decay.

Moreover, the asymmetry observed is as large as possible. In the electron angular distribution $I(\theta)=1+A\langle J_z\rangle/J(v/c)\cos\theta$ (where θ is the angle between the nuclear spin and electron momentum direction), the measured asymmetry parameter A is nearly equal to -1. This implies that the parity interference effects are about as large as they could be. So this first experiment on parity tells us not only that parity and charge conjugation are not conserved in beta decay but points to something even more drastic and significant.

TWO-COMPONENT THEORY OF THE NEUTRINO

When the experimental value of the asymmetry parameter ($A\cong-1$) in Co^{60} beta decay was made known to Lee and Yang, they immediately realized that here one had to consider an extremely simple and appealing theory of the neutrino.[3] This theory requires that the spin of a neutrino always be either parallel or antiparallel to its momentum and the helicity of an antineutrino be opposite to that of a neutrino. Incidentally, this theory requires also the rigorous masslessness of the particle. If there were any mass associated with the particle, the particle could therefore be at rest or with momentum $\mathbf{P}$ reversed in a certain frame of reference. In that case, it is rather meaningless to impose the necessary requirement of alignment of spin $\boldsymbol{\sigma}$ and momentum $\mathbf{P}$ for such a particle. From experimental evidence, the mass of the neutrino is indeed vanishingly small. The most sensitive method of estimating the mass of the neutrino is to investigate the slope of the upper end of a beta spectrum. The low-energy beta spectrum of H^3 has been investigated for this purpose by many laboratories.[4] The upper limit of the mass of the neutrino m_ν is around 250 ev ($\sim 1/2000\, m_e$) and the evidence is also not inconsistent with mass m_ν equal to zero.

All through the years, theoretical physicists had entertained the idea of associating such unique properties with a massless neutrino. It had to be abandoned lest the law of parity be violated. Clearly, the two-component theory of the neutrino violates P and C invariance separately. A left-handed neutrino looks into a mirror and finds a right-handed neutrino, but according to the two-component theory of the neutrino, there is no such possible state. The charge conjugate of a neutrino is an antineutrino; one behaves like a left

* Work partially supported by U. S. Atomic Energy Commission.

[1] H. Weyl, *Symmetry* (Princeton University Press, Princeton, New Jersey, 1952); E. P. Wigner, Proc. Am. Phil. Soc. **93**, 521 (1949); C. N. Yang, Science **127**, 565 (1958).

[2] Wu, Ambler, Hayward, Hoppes, and Hudson, Phys. Rev. **105**, 1413 (1957).

[3] T. D. Lee and C. N. Yang, Phys. Rev. **105**, 1671 (1957).

[4] Curan, Angus, and Cockcroft, Phil. Mag. **40**, 36, 53 (1949); G. C. Hanna and B. Pontecorvo, Phys. Rev. **75**, 983 (1949); Hamilton, Alford, and Gross, Phys. Rev. **83**, 215 (1951); and L. M. Langer and R. J. D. Moffat, Phys. Rev. **88**, 689 (1952).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

spinning particle, the other like a right spinning particle. This is a clear-cut violation of invariance of charge conjugation. In 1929, Weyl[5] proposed the mathematical possibility of such a particle, but it was discarded because it violated the law of parity and therefore could not be a physical reality! Landau[6] and Salam[7] reinvestigated this possibility shortly before the discovery of parity nonconservation. With the removal of both of these obstacles, interest in this theory was suddenly revived.

So in the first parity experiment of beta decay of polarized Co^{60}, not only was the screw sense in beta decay found, but even more dramatic was the realization of the possible existence of the two intrinsic opposite screws associated with the neutrino and antineutrino. The success of the two-component theory of the neutrino greatly facilitated our understanding of many phenomena in weak interactions.

LAW OF CONSERVATION OF LEPTONS

At this point, we should review another important conservation law known as the law of conservation of leptons which was suggested by Konopinski and Mahmoud[8] to explain the nonoccurrence of certain decay processes. The law states that if a leptonic number is assigned to each particle, then the sum of leptonic numbers must be conserved in all reactions. The assignments generally agreed upon are

lepton l=same (say +1)	for e^-, μ^-, ν
$l=-1$	for e^+, μ^+, $\bar{\nu}$
$l=0$	for π, γ, K and all heavy particles.

The ν or $\bar{\nu}$ expected in various decays are shown in the following equations:

$$n \rightarrow p+e^-+\bar{\nu} \qquad p \rightarrow n+e^++\nu \tag{1}$$

$$\pi^+ \rightarrow \mu^++\nu; \qquad \pi^- \rightarrow \mu^-+\bar{\nu} \tag{2}$$

$$\mu^+ \rightarrow e^++\nu+\bar{\nu}; \qquad \mu^- \rightarrow e^-+\nu+\bar{\nu} \tag{3}$$

$$\mu^-+p \rightarrow n+\nu; \qquad \mu^++n \rightarrow p+\bar{\nu} \tag{4}$$

$$K^+ \rightarrow \mu^++\nu; \qquad K^- \rightarrow \mu^-+\bar{\nu} \tag{5}$$

$$\pi^+ \rightarrow e^++\nu; \qquad \pi^- \rightarrow e^-+\bar{\nu}. \tag{6}$$

The confirmation of the assignments given by these decays should constitute a strong proof of the law of conservation of leptons. Particularly if the helicities of these two neutrinos are opposite, then the predictions of these decay processes become unique.

[5] The possibility of a two component relativistic theory of a spin 1/2 particle was first discussed by H. Weyl, Z. Physik **56**, 330 (1929). It was rejected on the ground of parity violation. See W. Pauli, *Handbuch der Physik* (Verlag Julius Springer, Berlin, 1933), Vol. 24, pp. 226–227.

[6] L. Landau, Nuclear Phys. **3**, 127 (1957).

[7] A. Salam, Nuovo cimento **5**, 299 (1957).

[8] E. Konopinski and H. M. Mahmoud, Phys. Rev. **92**, 1045 (1953); T. D. Lee and C. N. Yang, Phys. Rev. **105**, 1671 (1957).

EVIDENCE OF TWO-COMPONENT NEUTRINOS AND LEPTON CONSERVATION OBTAINED FROM PARITY EXPERIMENTS

I. From Nuclear Beta Decays

In the nuclear beta decay as shown in (1), an antineutrino is expected to be associated with β^- decay and a neutrino with β^+ decay. Furthermore, if the two-component theory of the neutrino is valid, then there are several unique phenomena predicted theoretically. The validity of the two-component theory and lepton conservation depends largely on the closeness of the agreement between the experimental evidence and the theoretical predictions. Suppose now we examine each phenomenon separately.

(A) *Beta Asymmetry from Polarized Nuclei*[2]

The beta asymmetry of allowed transitions from polarized nuclei can be expressed in the following form:

$$\begin{aligned} A=\Big\{&\lambda_{JJ'}\Big[\pm\mathrm{Re}(C_T{}^*C_{T'}-C_A{}^*C_{A'}) \\ &-\frac{Ze^2}{\hbar cp}\,\mathrm{Im}(C_T{}^*C_{A'}+C_{T'}{}^*C_A)\Big]|M_{GT}|^2 \\ &+\delta_{JJ'}\left(\frac{J}{J+1}\right)^{\frac{1}{2}}\Big[\mathrm{Re}(C_T{}^*C_{S'}+C_{T'}{}^*C_S-C_A{}^*C_{V'} \\ &-C_{A'}{}^*C_V)\pm\frac{Ze^2}{\hbar cp}\,\mathrm{Im}(C_A{}^*C_{S'}+C_{A'}{}^*C_S-C_T{}^*C_{V'} \\ &-C_{T'}{}^*C_V)\Big]|M_F|\cdot|M_{GT}|\Big\}\times\frac{2}{\xi(1+b/W)} \end{aligned} \tag{7}$$

$$\xi=(|C_S{}^2|+|C_V{}^2|+|C_{S'}{}^2|+|C_{V'}{}^2|)|M_F|^2 +(|C_T{}^2|+|C_A{}^2|+|C_{T'}{}^2|+|C_{A'}{}^2|)|M_{GT}|^2$$

$$\xi b=\pm 2\gamma\,\mathrm{Re}[(C_SC_V{}^*+C_S{}'C_{V'}{}^*)|M_F|^2 +(C_TC_A{}^*+C_T{}'C_{A'}{}^*)|M_{GT}|^2],$$

where C_i and C_i' are the even and odd beta coupling constants. This measurement gives unique information about the pure G-T interaction. Neglecting the imaginary term for the time being and setting the Fierz term $b/W=0$, then one can obtain a simple relation for "A" for the $J \rightarrow J'=J-1$ transition,

$$A(\beta_\pm)=\pm 2[\mathrm{Re}(C_T{}^*C_{T'}-C_A{}^*C_{A'})]/ |C_T{}^2|+|C_A{}^2|+|C_T{}'|^2+|C_A{}'^2|.$$

Furthermore, for the two-component neutrinos the even and odd beta coupling constants are related by $C_T=-C_{T'}$ (left-handed antineutrino) or $C_A=C_{A'}$ (right-handed antineutrino). Therefore, $A(\beta_\pm)=\mp 1$ for the above pure G-T transition. As we all know, the β^- asymmetry A of polarized Co^{60} where the spin and

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

parity change is $5^+ \rightarrow 4^+$ is indeed nearly -1 to within a systematic uncertainty not larger than 20%. This is the first strong evidence of the possible existence of the two-component neutrino. On the other hand, what evidence do we have of the β^+ asymmetry in a pure G-T transition of β^+ decay? Unfortunately, with the present technique of nuclear alignment, the nuclei which can be substantially polarized belong to a very exclusive club. In order to obtain the same type of information which can be extracted from polarized nuclei, we have to resort to an indirect method which is easier to apply although far less sensitive. This is shown in the following.

(B) β-γ (*Circular Polarization*) *Correlation*

It is quite obvious from the observed beta-asymmetry distribution of polarized nuclei that the β decay should leave the nucleus partly polarized with respect to the direction in which the β particle is detected. If a γ ray follows immediately after the β decay, it should have circular polarization proportional to the cosine of the angle between the β particle and γ ray. The circular polarization of the γ ray can be analyzed with a cylindrical electromagnet which could be magnetized to saturation either parallel or anti-parallel to the photon direction. The principle of this analysis is based on the existence of a spin dependent part of the Compton cross section of circularly polarized photons and is treated in detail by Gunst and Page.[9] The correlation for the most frequent decay sequence such as

$$J \xrightarrow[\beta \text{ transition}]{\text{allowed}} J' \xrightarrow{2^L \text{ pole } \gamma} J''=J'-L$$

can be expressed by

$$W(\theta)=1+A\frac{v}{c}\cos\theta,$$

where

$$A+\frac{1}{L+1}\Big\{\mu_{JJ'}\Big[\pm\mathrm{Re}(C_T{}^*C_T'-C_A{}^*C_A')-\frac{\dot{Z}e^2}{\hbar cp}\,\mathrm{Im}(C_T{}^*C_A'+C_T'{}^*C_A)\Big]|M_{GT}|^2+\delta_{JJ'}\Big(\frac{J+1}{1}\Big)^{\frac{1}{2}}\Big[\mathrm{Re}(C_T{}^*C_S'+C_T'{}^*C_S-C_A{}^*C_V'-C_A'{}^*C_V)\pm\frac{Ze^2}{\hbar cp}\,\mathrm{Im}(C_A{}^*C_S'+C_A'{}^*C_S-C_T{}^*C_V'-C_T'{}^*C_V)\Big]|M_F|\cdot|M_{GT}|\Big\}\frac{2}{\xi(1+b/W)}. \quad (8)$$

Schopper[10] first applied this β-γ (circular) correlation method to Co^{60} and obtained the parameter $A\cong-\frac{1}{3}$. This is in excellent agreement with the conclusion derived from the polarized Co^{60} experiment, that is, $C_T=-C_T'$ or $C_A=C_A'$. When this method was applied to Na^{22}, which is a pure G-T positron emitter ($3^+ \rightarrow 2^+$), the sign of the circular polarization was found to be opposite to that of the electron emitter Co^{60} as theoretically predicted. However, none of these experiments[11] has an accuracy of much better than 20%.

(C) *Longitudinal Polarization of β Rays*[12]

Since parity is not conserved, the pseudoscalar term $(\sigma\cdot p_e)$ formed from the measured spin and momentum vector of electrons may occur. In other words, the decay electrons from unpolarized nuclei can be longitudinally polarized. The general expression of the $\beta^\pm$ polarization is

$$P(\beta^\mp)=\pm\frac{v}{c}\frac{2\,\mathrm{Re}[(C_SC_S'{}^*-C_VC_V'{}^*)|M_F|^2+(C_TC_T'{}^*-C_AC_A'{}^*)|M_{GT}|^2}{(|C_S{}^2|+|C_V{}^2|+|C_S'{}^2|+|C_V'{}^2|)|M_F|^2+(|C_T{}^2|+|C_A{}^2|+|C_A'{}^2|+|C_T'{}^2|)|M_{GT}|^2}. \quad (9)$$

If we assign $C_S=-C_S'$; $C_T=-C_T'$ (for left-handed antineutrino) or $C_V=C_V'$; $C_A=C_A'$ (for right-handed antineutrino) then $P(\beta^\mp)=\mp v/c$. First to observe this polarization was the Frauenfelder group. There are, at present, three major methods being used for the determination of electron polarization. They are: (1) Coulomb scattering from heavy nuclei (Mott scattering); (2) circular polarization of forward Bremsstrahlung or annihilation radiation; and (3) free electron-electron scattering (Møller or Bhabha scattering). The results are startling and simple, and all agree that β^- particles emitted in radioactive decay behave like left-handed screws and all β^+ particles behave like right-handed screws. For relativistic energy $v/c\cong1$, we have practically completely polarized elec-

[9] S. B. Gunst and L. A. Page, Phys. Rev. **92**, 970 (1953); Wheathly, Huiskamp, Diddens, Steenland, and Tolhoek, Physica **2**, 841 (1955). W. J. Huiskamp, thesis, Leiden, 1958; and H. Schopper, Nuclear Instr. **3**, 158 (1958).

[10] H. Schopper, Phil. Mag. **2**, 710 (1957).

[11] F. Boehm and A. N. Wapstra, Phys. Rev. **106**, 1364 (1957); P. Debrunner and W. Kündig, Helv. Phys. Acta **30**, 261 (1957).

[12] *Proceedings of the Rehovoth Conference on Nuclear Structure*, (North-Holland Publishing Company, Amsterdam, 1958). (a) Mott Scattering: H. A. Tolhoek, Revs. Modern Phys. **28**, 177 (1956); Frauenfelder, Bobone, von Goeler, Levine, Lewis, Peacock, Rossi, and De Pasquali, Phys. Rev. **106**, 386 (1957); Alikhanov, Yeliseyev, Linbimov, and Ershler, Cavanagh, Turner, Coleman, Gard, and Ridley, Phil. Mag. **2**, 1105 (1957); H. de Waard and O. J. Poppema, Physica **23**, 597 (1957); de-Shalit, Kuperman, Lipkin, and Rothem, Phys. Rev. **107**, 1459 (1957); and Langevin-Joliot, Marty, and Sergent, Compt. rend. **244**, 3142 (1957).
(b) Circular polarization of bremsstrahlung and annihilation radiation: K. M. McVoy, Phys. Rev. **106**, 828 (1957); *ibid.* **110**, 1484 (1958); Goldhaber, Grodzins, and Sunyar, Phys. Rev. **106**,

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

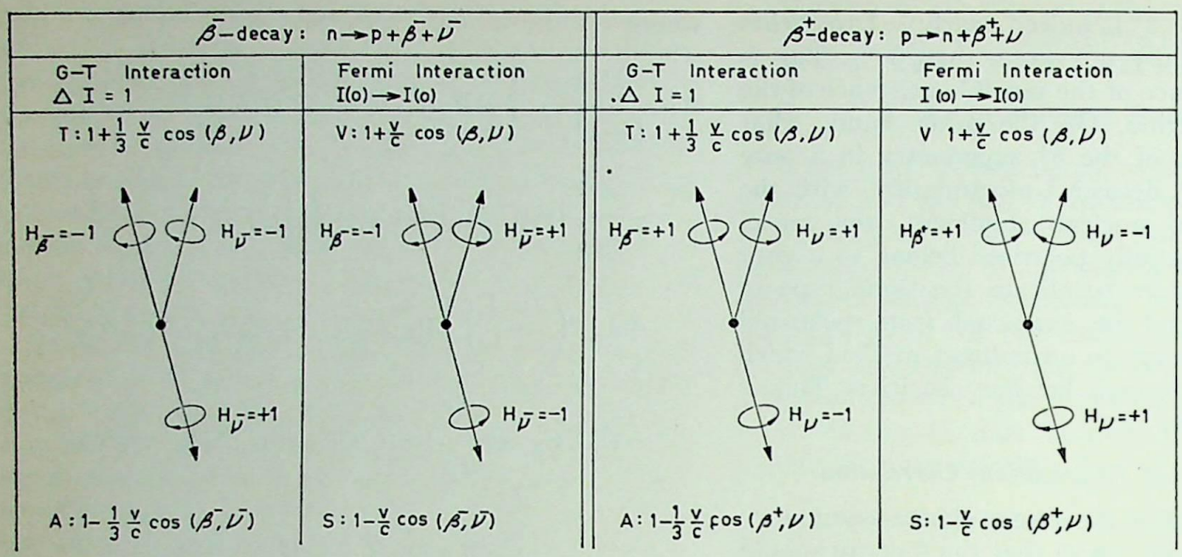


FIG. 1. Correlations of neutrino helicities and β interactions.

tron or positron beams. We have been working with the polarized beam of β particles for the past sixty years, yet we were completely unaware of it because of the faultiness of our left-right symmetry conception. However, many systematic uncertainties in the measurements, such as the backscattering effect, depolarization effect, instrumental asymmetry, and the screen correction factor, etc., are rather difficult to assess or to estimate. Probably it is fair to say that in the high-energy region of $v/c \geq 0.6$, the polarization is nearly v/c with an accuracy of not better than 10%. Below $v/c=0.6$ very few results have been reported and the picture in that region is not clear.[13]

(D) *Correlations between the Helicities of Leptons and Beta Interactions*

The helicity of the neutrino can be correlated to the electron polarization and the form of beta interaction as shown in Fig. 1.

In a G-T β^- interaction, the angular momentum carried away by the leptons is one unit. In the tensor interaction, both leptons are emitted preferentially in the same direction. Since the electron is found to possess negative helicity (left-handed screw), the antineutrino must have the same helicity. On the other hand, in the axial vector interaction, the electron and the antineutrino are emitted preferentially in opposite directions; so positive helicity is predicted for the antineutrino and negative helicity for the neutrino. The helicity of the neutrino[14] observed in the electron capture process of Eu^{152*} is indeed negative and therefore supports the axial-vector interaction. In a similar manner, one can figure out the helicity of the antineutrino for scalar and vector interactions. The relations between the helicities of various leptons in different beta interactions are shown in Table I. This table also points out two more significant conclusions: First, the much used Fierz interference term between S and V expressed as $(C_S C_V{}^* + C_S{}' C_V{}'^*)$ and A and T such as $(C_T C_A{}^* + C_T{}' C_A{}'^*)$ now automatically vanishes because the neutrino or antineutrino associated with S and T or V and A has opposite helicity. Therefore, no interference occurs. Secondly, for the same reason, combinations of (V and T) and (S and A) will not result in interference between Fermi and Gamow-Teller terms.

Many beautiful β-γ (circular polarization) correlation studies[15] on the mixed (G-T and Fermi) transitions such as Sc^{46}, Au^{198}, etc., showed nearly maximum amount of interference between G-T and Fermi interactions. This evidence strongly ruled out a pure VT or SA combination. Incidentally, this conclusion had been reached also by Mahmoud and Konopinski[16] by the study of the first forbidden beta spectra alone in the pre-parity era.

TABLE I. Helicities of ν, $\bar{\nu}$, β^-, and β^+. Helicity: $\boldsymbol{\sigma}\cdot\mathbf{p}/|p|$.

Helicities	ν	$\bar{\nu}$	β^-	β^+
S, T	$+1$	-1	$-\frac{v}{c}$	$+\frac{v}{c}$
V, A	-1	$+1$	$-\frac{v}{c}$	$+\frac{v}{c}$

826 (1957); Deutsch, Gittelman, Bauer, Grodzins, and Sunyar, Phys. Rev. **107**, 1733 (1957); and Boehm, Novey, Barnes, and Stech, Phys. Rev. **108**, 1497 (1957).

(c) Møller Scattering: C. Møller, Ann. Physik **14**, 531 (1932); L. A. Page, Phys. Rev. **106**, 394 (1957); A. M. Bincer, Phys. Rev. **107**, 1434, 1469 (1957); G. W. Ford and C. J. Mullin, Phys. Rev. **108**, 477 (1957); Frauenfelder, Hanson, Levine, Rossi, and De Pasquali, Phys. Rev. **107**, 643, 909, 910 (1957); and Benczer-Koller, Schwarzschild, Vise, and Wu, Phys. Rev. **109**, 193 (1958).

[13] Tumer, Gard and Cavanagh, Paper S1, Bulletin of Conference on Weak Interactions, Gatlinburg, Tennessee, 1957.

[14] Goldhaber, Grodzins, and Sunyar, Phys. Rev. **109**, 1015 (1958).

[15] F. Boehm and A. H. Wapstra, Phys. Rev. **109**, 456 (1958); Lundby, Patro, and Stroot, Nuovo cimento **6**, 745 (1957); Steffen and Alexander, Proceedings of Rehovoth Conference (1957); and W. Güngst and H. Schopper, Z. Naturforsch **13a**, 505 (1958).

[16] H. M. Mahmoud and E. J. Konopinski, Phys. Rev. **88**, 1266 (1952).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

(E) *Inverse Beta Decays and the Double Beta Decays*

There are also three experiments, initiated long before the parity crisis, which originally had no bearing on the parity question. However, it turns out now that they have important implications for the two-component neutrino and the conservation of leptons. As a matter of fact, a voice of dissension fron these investigations could cause serious troubles.

1. Capture cross section for the antineutrinos in the Cowan and Reines Experiments.[17]—$\bar{\nu}+p \rightarrow n+\beta^+$. For two-component neutrinos, the outgoing neutrinos from the reactor have only one spin state instead of the usual two. By a detailed balancing method, the absorption cross section will be twice as great as the old one. The latest experimental results[18] are no longer in disagreement with the two-component theory of the neutrino. The cross section per fission $\bar{\nu}$ (assuming 6.1 $\bar{\nu}$ per fission) for the inverse beta decay of the proton is

$$\sigma_{\text{exp}}=(11\pm4)\times10^{-44}\ \text{cm}^2,$$

which is comparable to the theoretical cross section calculated for the two-component theory of the neutrino and based on the antineutrino spectrum from the fission fragments.

$\sigma_{\text{theo}}=9.5\times10^{-44}$ Reines *et al.*[18]
$=12\times10^{-44}$ Muehlhause and Oleksa[18]
$=15\times10^{-44}$ King and Perkins.[18]

2. Capture process for the neutrino.—In the Davis experiment[19] $\nu+\text{Cl}^{37} \rightarrow \text{A}^{37}+\beta^-$. Here a neutrino, not an antineutrino, is required for the absorption process. The intensive flux of antineutrinos pouring out of a power nuclear reactor does not provide the right kind of neutrino for this process and therefore no A^{37} activity should be attributed to the reactor neutrinos. The very small A^{37} counts (0.3±3.4 counts per day) which Davis observed in his one thousand gallons of CCl_4 are equivalent to a cross section for neutrino capture of $(0.1\pm0.6)\times10^{-45}$ cm^2. This could be accounted for by the muon activation from the cosmic radiation, and no evidence for a positive effect from the reactor neutrinos exists at the present time. At least, there is no dissension about the two-component theory of the neutrino from the capture process of the neutrino.

3. Double beta-decay.[20a]—The theoretically calculated rates of double beta-decay depend on many assumptions concerning the properties of the neutrinos and whether one or both kinds of neutrinos are emitted in β^- decay. The existence of the conservation law of leptons, together with the two-component neutrino, limits the possible alternatives and demands that the rate of double beta decay be its minimum value. Experimentally no single electron line of discrete energy equal to the sum of the two disintegration energies has ever been observed. The observed upper limit of the rate of double beta decay also strongly rejects the short lifetime based on Majorana neutrinos and favors the longer time of Dirac neutrinos. For example, the predicted half-life of double beta decay of Ca^{48} (4.3±0.1 Mev) is 4×10^{15} yr for no neutrino emissions and 4×10^{18} yr for Dirac neutrinos. The observed decay rate of Ca^{48} is $>2\times10^{18}$ [20b] or 6×10^{18} yr.[20b] In the case of Nd^{150} (3.7±0.1 Mev), the observed decay rate is $>2\times10^{18}$ yr[20c] in comparison with that predicted for Dirac neutrinos of 2×10^{18} yr. However, it is desirable to increase the sensitivity of the detector so one can actually observe the Dirac type double beta decay.

So, all in all, the experimental evidence in favor of two-component neutrinos and the conservation law of leptons is overwhelming. Not even a faint voice of dissension has been raised. On the other hand, neither can a great accuracy be claimed. A greatly improved accuracy to strengthen its ultimate validity is much to be desired.

II. From $\pi-\mu-e$ Decays

Here we review briefly how successfully the two-component theory of the neutrino and the conservation of leptons can interpret the observed phenomena in $\pi-\mu-e$ decays. Particularly, we show that the observed helicities of the neutrino and the antineutrino in $\pi-\mu-e$ decays are the same as those found in nuclear beta decays.

(A) *Polarization of Muons and Electrons*

Consider the decay of a positive pion in its own rest frame [Fig. 2(a)]. Since π^+ is spinless, the neutrino angular momentum (its spin 1/2) about its direction of motion must be balanced by the muon angular momentum about its direction of motion. Now the neutrino in nuclear beta decay is found to possess a negative helicity, so negative helicity is predicted for the μ^+. So far no direct determination of μ^+ helicity has been reported.

However, the information may be inferred from the helicity of the decay electrons within the two-component theory. In the extreme case [Fig. 2(a)] when the neutrino and the antineutrino go in the same direction, the electron which goes in the opposite direction has to carry away the angular momentum of the muon. If μ^+ has a negative helicity, the decay e^+ must possess positive helicity. The experimental results on the polarization of the negative and positive electrons from

[17] F. Raines and C. L. Cowan, Phys. Rev. **92**, 830 (1953); Cowan, Reines, Harrison, Kruse, and McGuire, Science **124**, No. 3212, 103 (1956).

[18] F. Reines *et al.*, *Second Atoms for Peace Conference, Geneva*, 1958, Paper No. 1026.

[19] R. Davis, Phys. Rev. **97**, 766 (1955); Bull. Am. Phys. Soc. Ser. II, 1, 219 (1956); and private communication (1957).

[20] (a) H. Primakoff and S. P. Rosen, "Double beta decay," Washington University, 1958 (unpublished). (b) Ca^{48}: M. Awschalom, Phys. Rev. **101**, 1041 (1956); Dobrokhotov, Lazarenko, and Luk'yanov, *CERN High Energy Conference, 1958*. Nd^{150}: Cowan, Harrison, Langer, and Reines, Nuovo cimento **3**, 649 (1956).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

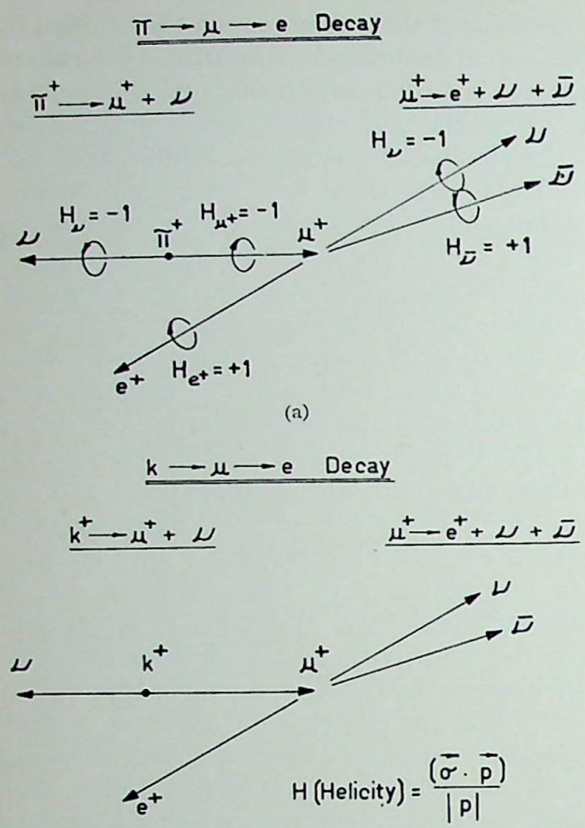


FIG. 2.

muon decays obtained by measuring the circular polarization of the bremsstrahlung and annihilation radiation[21] or by the Møller scattering method[22] conclusively show positive helicity for e^+ and negative helicity for e^-. Here the results on the helicities of the neutrinos in nuclear beta-decays and from $\pi-\mu-e$ decays are in excellent accord.

(B) *Michel Parameter "ρ"*

The shape of the energy spectrum of e^+ from μ^+ decay is characterized by a single parameter "ρ" discussed by Michel. In the two-component theory of the neutrino, ρ must equal zero if the electron is accompanied by two neutrinos or two antineutrinos, and equal 3/4 if one neutrino and one antineutrino are emitted. The present experimental values[23] of ρ vary from 0.68 to 0.72, which is close to 0.75, for one neutrino and one antineutrino emission, and confidently reject the assumption of two neutrinos or two antineutrinos in muon decay. The ultimate determination of the value of ρ is highly desirable because of its significant theoretical implications.

(C) *Energy Dependence of the Asymmetry Coefficient*[24]

The two-component theory of the neutrino predicts closely the energy dependence of the asymmetry coefficient of electron distribution in muon decay. In the high-energy end of the spectrum, the data are in good agreement with the theory. At the low-energy region, measurements are difficult and the accuracy has been poor. However, as a whole, the energy dependence of the asymmetry coefficient in μ^- decay agrees with the prediction of the two-component theory.

III. Muon Capture

The process of muon capture has been studied only in complex nuclei and very little information is available in the literature. At present one side of the Puppi triangle, which represents the muon capture process, is attracting much attention. Before the conference is over we will hear several interesting papers[25] and communications[25] on this topic.

IV. $K-\mu-e$ Decays[26]

The $K-\mu-e$ decay is analogous to $\pi-\mu-e$ decay [Fig. 2(b)]. If the conservation law of leptons holds, the same kind of neutrinos are expected in these two parallel cases. This should, therefore, result in the same asymmetry distribution of electrons with respect to the direction of μ motion. The observance of this predicted asymmetry in $K-\mu-e$ decay further strengthens the validity of lepton conservation.

V. $\pi-e$ Decays[27]

The existence of $\pi-e$ decays is finally confirmed by several laboratories, and the measured ratio R of $\pi\rightarrow e$ to $\pi\rightarrow\mu+\nu$ decay is very close to that predicted by a simple calculation which gives $R=(\pi-e)/(\pi-\mu)=1.36\times10^{-4}$. It is interesting to note that the helicity of the electron from $\pi-e$ decay should be opposite to that from $\mu-e$ decays.

So there is overwhelming evidence in favor of the two-component theory of the neutrino and the conservation of leptons in all the beta-decay phenomena. To establish firmly the validity of these two assumptions, more precision measurements are required.

INTERACTIONS RESPONSIBLE FOR BETA DECAY

I. From Classical Beta-Decay Experiments

Prior to the discovery of parity nonconservation in beta decay, the (S,T) combination had been the

[21] Culligan, Frank, Holt, Kluyver, and Massam, Nature **80**, 751 (1957); K. Crowe, Washington Meeting, American Physical Society (1958).

[22] H. Anderson, *Proceedings of the CERN Conference on High Energy Physics*, 1958.

[23] ρ value: 0.68±0.02, K. M. Crowe, Bull. Am. Phys. Soc. Ser. II, **2**, 234 (1957). 0.68±0.09, Sargent, Rinehart, Lederman, and Rogers, Phys. Rev. **99**, 885 (1955). 0.72±0.05, W. Dudziak and R. Sagane, Rochester Conference on High Energy Physics (1957). 0.67±0.05, L. Rosenson, thesis, Chicago University (1957).

[24] *Proceedings of the CERN Conference on High Energy Physics*, 1958.

[25] Proceedings and Bulletin of Conference on Weak Interactions, 1958, Sessions F and G.

[26] Coombes, Cork, Galbraith, Lambertson, and Wenzel, Phys. Rev. **108**, 1348 (1957).

[27] T. Fazzini, Fidecaro, Merrison, Paul, and Tollestrup, Phys. Rev. Letters **1**, 247 (1958); G. Impeduglia *et al.*, Phys. Rev. Letters **1**, 249 (1958); and Proceedings of Conference on Weak Interactions, 1958.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

favorite choice based mainly on $He^6(\beta\text{-}\nu)$ angular correlation results.[28] In fact, the $(\beta\text{-}\nu)$ angular correlation was the only means used to investigate the type of beta interaction in those days. But this type of experiment was known to be difficult. Wu and Schwarzschild[29] made a detailed examination of the old He^6 experiment and pointed out that the effective volume of the He^6 source in the hole of the pumping diaphragm was not correctly taken into account. Had this been done properly, the results of He^6 would not then have implied the tensor interaction. However, in spite of its many limitations, $\beta\text{-}\nu$ correlation is still an effective and powerful method in yielding information on the beta interactions. The first sign of warning against the (S,T) combination in beta interaction came in May of 1957 when Herrmannsfeldt *et al.*[30] published their $(\beta\text{-}\nu)$ correlation results on A^{35}, which decays mainly through Fermi interaction, and the results strongly supported the vector interaction instead of scalar as was once believed.

Recently,[31] this group has remeasured the $(\beta\text{-}\nu)$ correlation in He^6 with the same apparatus and obtained the correlation coefficient $\lambda=-0.39\pm0.02$, which certainly favors axial vector in He^6.

Meantime the $\beta\text{-}\nu$ angular correlation in

$$Li^8 \xrightarrow{\beta} Be^8 \longrightarrow 2\alpha$$

was reinvestigated with greatly improved technique at California Institute of Technology[32] and at Heidelberg University.[33] They measured either the angular distribution of the two α particles or the α spectra in coincidence with the β particles emitted at either 180° or 90° from the α direction. The results from Heidelberg strongly favor the axial vector, and the California Institute of Technology results are accurate enough to put the upper limit of tensor mixture in G-T interaction to less than 10%.

Thus the (V,A) combination is strongly favored from $(\beta\text{-}\nu)$ correlation experiments. Let us now look into the information on beta interactions obtained from parity experiments.

II. From Parity Experiments

(A) *Electron Capture Process in* Eu^{152*} [34]

In an electron capture process, a neutrino and the recoil nucleus are emitted in opposite directions $e^- + p \rightarrow n + \nu$. If the capture process is followed by the emission of a gamma ray and the spin and parity changes are favorable as shown in the following decay process

$$A(0^-) \xrightarrow{e^- \text{ capture}} B^*(1^-) \xrightarrow{\gamma \text{ ray}} B(0^+)$$

then, by applying the conservation laws of momentum and angular momentum, one can deduce a simple correlation that the helicity of the downward gamma ray will be the same as that of the upward neutrino. So the problem of determining the neutrino helicity becomes that of measuring the circular polarization of the gamma ray. However, to select only those downward gamma rays following the emission of the upward neutrinos, many conditions must be fulfilled. First, the gamma ray must have an energy comparable to that of the neutrino and the lifetime of the excited level B^* must be very short ($\sim 10^{-14}$ sec) in order to permit the use of solid material. Even then one has to detect only the resonantly scattered gamma ray.

The requirements were indeed strict, but the radioisotope Eu^{152*} seemed heaven-sent to do this job. Goldhaber, Grodzins, and Sunyar knew of this radioisotope Eu^{152*} from their previous investigations and it fulfills all the requirements stated above. By measuring the circular polarization of the gamma rays from Eu^{152*} which are resonantly scattered by Sm, they found that the helicity of the gamma ray is *negative*! ($H=-0.67\pm0.10$). From this result one concludes that the helicity of the neutrino in electron capture is negative and therefore the Gamow-Teller interaction in electron capture is dominantly "A" and not "T."

(B) *Beta Decay of Polarized Neutrons*[35]

Meanwhile, at the Argonne National Laboratory, a highly polarized neutron beam had been successfully completed since 1957. Burgy *et al.*[35] had been measuring two asymmetry coefficients from the beta decays of polarized neutrons. One is the coefficient "A," the β^- distribution asymmetry ($\langle J\rangle \cdot p_{e^-}$); the other is the coefficient "B," the $\langle\nu\rangle$ distribution asymmetry ($\langle J\rangle \cdot p_{\langle\nu\rangle}$). From the value $A=-0.11\pm0.02$ and $B=0.88\pm0.15$, they concluded that the interaction in β^- decay is dominantly V and A with opposite phase relations $(V-A)$.

Look back at the history of the theory of beta decay. It has been filled with surprises and excitement. Now, after a period of nearly sixty years of continuous investigation, finally along comes the nonconservation of parity. These two (classical beta theory and nonconservation of parity) joined forces in reaching the conclusion on the beta interaction.

However, we must be cautioned here that whether there is any small mixture of (S,T) in beta interaction is still unknown and is waiting to be proved.

---

[28] B. N. Rustad and S. Ruby, Phys. Rev. **97**, 991 (1955).
[29] C. S. Wu and A. Schwarzschild, Columbia University Report CU-173.
[30] Hermannsfeldt, Maxson, Stähelin, and Allen, Phys. Rev. **107**, 641 (1957).
[31] Herrmannsfeldt, Burman, Stähelin, Allen, and Braid, Phys. Rev. Letters **1**, 61 (1958).
[32] Lauritsen, Barnes, Fowler, and Lauritsen, Phys. Rev. Letters **1**, 326 (1958); Barnes, Fowler, Greenstein, Lauritsen, and Nordberg, Phys. Rev. Letters **1**, 328 (1958).
[33] Lauterjung, Schimmer, and Maier-Leibnitz, Z. Physik **150**, 657 (1958).
[34] Goldhaber, Grodzins, and Sunyar, Phys. Rev. **109**, 1015 (1958).
[35] Burgy, Krohn, Novey, Ringo, and Telegdi, Phys. Rev. **110**, 1214 (1958).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

III. Universal Fermi Interaction

The great similarity in the strength of the coupling constants in beta decay, μ decay and μ capture suggests that the interaction forms of the three decay processes may also be the same. When beta-decay coupling was concluded to be (S,T) and μ-decay coupling was deduced to be dominantly V and A from the negative sign of the asymmetry coefficient, this possibility of a universal Fermi interaction was naturally ruled out. Now that the interaction (V,A) has replaced (S,T) in beta decay, the situation is quite altered. This is particularly intriguing because this specific form of (V and A) has been prophesied independently by Marshak and Sudershan[36] and also by Feynmann and Gell-Mann[37] by different deductions. The interactions which are responsible for the muon capture process are now under intensive investigation.

CONCLUSION

We have come a long way since the overthrow of the law of parity in weak interactions. A great advance has been made in the theory of the neutrino. We have two-component neutrinos and the law of conservation of leptons. A vast amount of experimental data is accounted for by the $V-A$ theory, and therefore the idea of the universal $(V-A)$ Fermi interaction is again made acceptable.

However, there are still many questions unanswered. Is there any connection between the beta interactions and the gravitational forces? Why should the ratio of the coupling constants between G-T and Fermi interactions $|C_{GT}/C_F|^2$ be what was observed? Are there any possible theoretical explanations for this? The ratio of $|C_{GT}/C_F|^2$ calculated from the recent neutron half-life[38] of (11.7±0.3) min and the O^{14} ft value[39] is 1.42±0.08. The ratio of $|C_{GT}/C_F|^2$ from the asymmetry distribution of electrons from polarized neutrons is 1.56±0.14.[35] On the other hand, the ratio of $|C_{GT}/C_F|^2$ obtained from the analysis of the ft values of beta decays of mirror nuclei by the "B-X" diagram is only 1.16±0.05.[40] The last method relies largely on the evaluation of the matrix elements. Would this apparent discrepancy eventually show us a way to improve our evaluation of the matrix elements? We are all eagerly anticipating hearing many interesting discussions on these subjects at this conference.

[36] G. Sudarshan and R. Marshak, Padua-Venice International Conference, 1957, Phys. Rev. **109**, 1860 (1958).
[37] R. Feynmann and M. Gell-Mann, Phys. Rev. **109**, 193 (1958).
[38] Sosnovskij, Spivak, Prokofiev, Kutikov, and Dobrynin, *Proceedings of the CERN Conference on High Energy Physics*, 1958.
[39] J. B. Gerhart, Phys. Rev. **109**, 897 (1958).
[40] O. C. Kistner and B. M. Rustad, Bulletin of Conference on Weak Interactions, 1958, Paper E5.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

Determination of the Beta-Decay Interaction from Electron-Neutrino Angular Correlation Experiments

JAMES S. ALLEN

Department of Physics, University of Illinois, Urbana, Illinois

INTRODUCTION

EXPERIMENTAL determination of the specific form of the interactions responsible for nuclear beta decay has remained a critical problem ever since the original formulation of the theory by Fermi.[1] Bloch and Moller[2] probably were the first to point out that the electron-neutrino angular correlation in nuclear beta decay depends on the type of light-particle-neutron interaction assumed in the Fermi theory. According to present ideas, a linear combination of five relativistically invariant expressions can be chosen for the interaction Hamiltonian. The individual interactions are the scalar, vector, tensor, axial-vector, and pseudoscalar interactions, denoted respectively S, V, T, A, and P. In the case of allowed transitions, the pseudoscalr interaction usually is small compared to the other forms and can be omitted.

During the past ten years we have been conducting a series of electron-neutrino angular correlation experiments on allowed beta decays in an attempt to obtain an unambiguous determination of the form of the beta-decay interaction. In general, these experiments have tended toward a procedure which requires only a direct measurement of the shape of the energy spectrum of the recoiling nuclei. This method has eliminated the difficulties inherent in the experiments which require an explicit measurement of either the angle between the electron and recoiling nucleus or the energy of the electron. Our present method does not require the detection of the recoiling nucleus and the decay electron in coincidence and, consequently, the counting rate is not limited by a decreasing true to chance rate.

This program finally has been completed with a series of measurements of the angular correlations in the allowed beta decay of He^6, Ne^{19}, Ne^{23} and A^{35}. Since the entire set of measurements was carried out in one spectrometer, the experimental results may be compared directly. The results prove that the dominant interactions are, vector for the Fermi type of decay, and axial-vector for the Gamow-Teller decays.

THEORY

In the usual theory of allowed beta decay the transition probability for the emission of an electron-neutrino pair is

$$P(W,\theta_{ev})\mp dW d\Omega_e=\frac{1}{(2\pi)^4}F(\pm Z,W)pW(W_0-W)^2\xi \times\left[1+\frac{b}{W}+\frac{\lambda p}{W}\cos\theta_{ev}\right]dW d\Omega_e, \quad (1)$$

where W and p, respectively, are the total energy and momentum of the electron. The angular correlation coefficient, expressed in terms of the coupling constants and nuclear matrix elements, is

$$\lambda\xi=(|C_V|^2+|C_V'|^2-|C_S|^2-|C_S'|^2)|\langle 1\rangle|^2 +\tfrac{1}{3}(|C_T|^2+|C_T'|^2-|C_A|^2-|C_A'|^2)|\langle\sigma\rangle|^2, \quad (2)$$

where

$$\xi=(|C_V|^2+|C_V'|^2+|C_S|^2+|C_S'|^2)|\langle 1\rangle|^2 +(|C_T|^2+|C_T'|^2+|C_A|^2+|C_A'|^2)|\langle\sigma\rangle|^2. \quad (3)$$

There is also a small additional term in λ which vanishes if time reversal is invariant. This term also is identically zero with the current choice of coupling constants, $C_{VA}=+C_{V'A'}$ and $C_{TS}=-C_{T'S'}$. It is interesting to note that the Fierz interference term b/W also is identically zero with this choice of coupling constants.

It is readily seen from (2) that λ assumes the values $+1$, -1, $+\frac{1}{3}$, and $-\frac{1}{3}$ for the pure V, S, T, and A interactions, respectively. Also, λ does not contain terms due to interference between the parity conserving and parity nonconserving interactions. Therefore, electron-neutrino angular correlation experiments of the type discussed here are not able to separate the primed and unprimed coupling constants. In other words, these experiments cannot show the degree of mixing of the parity conserving and nonconserving interactions.

The energy distribution function for the decay electrons is obtained by integrating $\cos\theta_{ev}$ in (1) over all values of θ_{ev}. Since $\cos\theta_{ev}$ is an odd function of the angle, the integral is zero and consequently the shape of the energy spectrum does not depend on λ. However, when the distribution in energy of the recoiling nuclei is measured, the integration is over values of θ_{ev} consistent with conservation of energy and linear momentum in the emission of the electron-neutrino pair. In this case the term containing λ does not vanish and therefore the shape of the energy spectrum depends on the form of the angular correlation.

In our most recent experiments we have measured

[1] E. Fermi, Z. Physik 88, 161 (1934).
[2] F. Bloch and C. Moller, Nature 136, 912 (1935).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

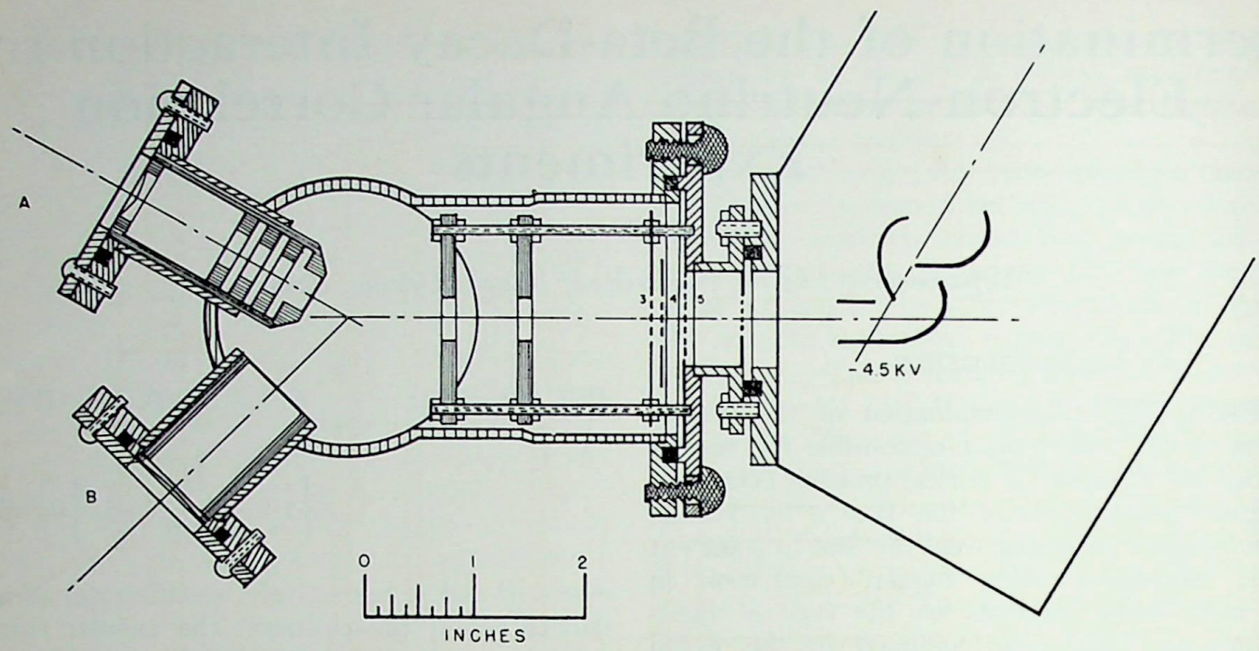


FIG. 1. The apparatus used for the first He^6 neutrino experiment. The recoil ions were counted by the electron multiplier and the electrons by a Geiger counter at window A or B. The energy spectrum of the recoil ions was measured by the retarding potential method.

the shape of the energy spectra of the recoiling nuclei and have deduced values of λ from comparisons of the experimental data with the theoretical distribution functions. The experimentally determined angular correlations have established the form of the dominant beta-decay interaction.

EARLY RECOIL EXPERIMENTS

The beta decay of He^6 can be represented by

$$He^6 \rightarrow Li^6 + e^- + \bar{\nu}, \quad \log ft = 2.91, \qquad (4)$$

with a spin change of $J=0$ to $J=1$, no change of nuclear parity. The maximum kinetic energy of the recoiling Li^6 ion is 1405 ev. The beta-decay process produces a singly charged positive ion. However, in about 10% of the decays, doubly charged ions are present as a result of ionization caused by the sudden change in the nuclear charge. The small value of ft indicates that the He^6 transition is allowed, and the spin change shows that this is a pure Gamow-Teller transition. The angular correlation factor will be either $+\frac{1}{3}$ for the tensor or $-\frac{1}{3}$ for the axial-vector interaction.

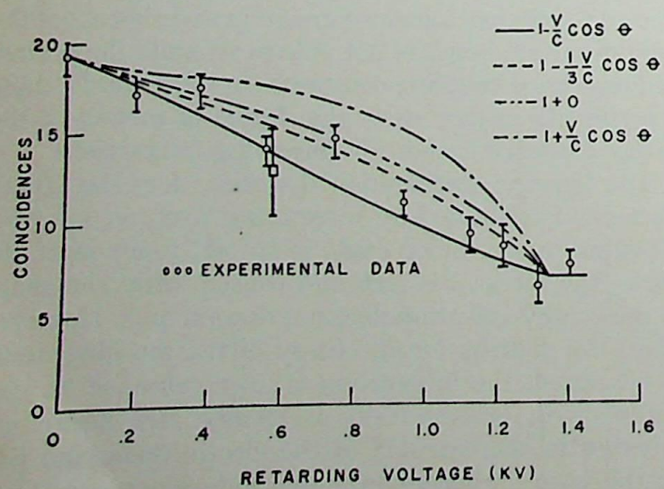


FIG. 2. Experimental and theoretical retarding potential curves for the recoil ions from the decay of He^6. The angle between the direction of emission of the electrons and the recoil ions was $162° \pm 8°$.

The first He^6 experiment was carried out by Allen, Paneth, and Morrish.[3] An attempt was made to obtain the recoil energy spectrum of the ions in coincidence with electrons emitted at angles of 135° and 162° with respect to the direction of the ions. A schematic of the apparatus is shown in Fig. 1. A retarding potential method was used for the analysis of the kinetic energy of the ions. The retarding voltage was applied to grid 4 located between the grounded grids 3 and 5.

The experimental data obtained with the Geiger counter at position A are shown in Fig. 2. Although the statistical accuracy is poor, most of the experimental points lie below the $\lambda=0$ curve and therefore suggest the axial-vector interaction.

The next experiment in our series represented an attempt to study the He^6 decay using a scintillation counter instead of a Geiger counter. A schematic diagram of the apparatus is shown in Fig. 3. The effective source volume was cylindrical in shape and was defined at one end by the scintillator and at the other end by the sliding gate A. This gate had an open and also a closed aperture. An aluminum foil covering the closed aperture was thick enough to stop the recoil ions, but also sufficiently thin to transmit the decay electrons without appreciable scattering or loss of energy. The difference between the counting rates observed with the gate open and closed represents the net rate produced by disintegrations occurring in the

[3] Allen, Paneth, and Morrish, Phys. Rev. 75, 570 (1949).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

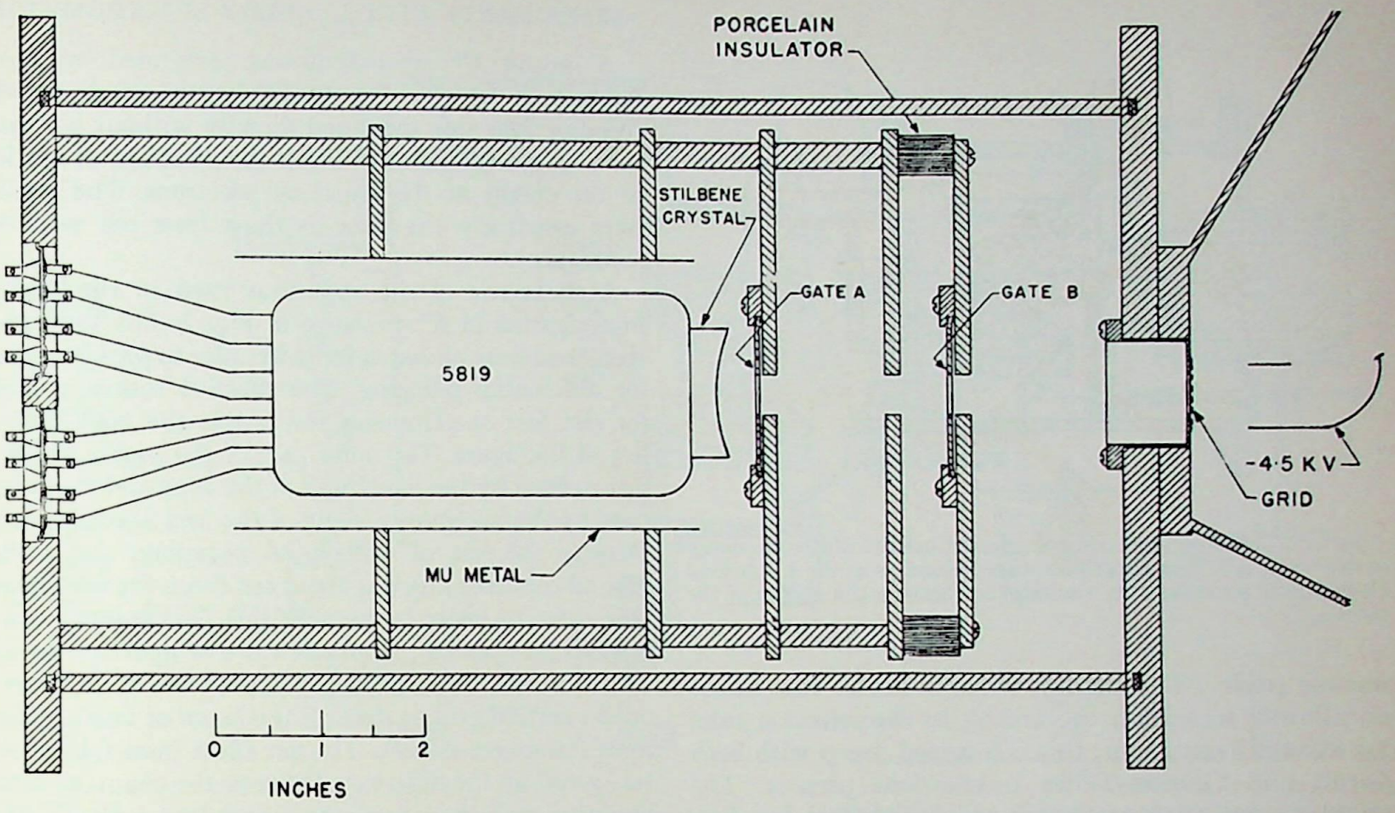


FIG. 3. The apparatus used for the second He^6 neutrino experiment. The effective source volume was cylindrical in shape and was defined at one end by the curved surface of the stillbene crystal and at the other end by the gate A. The velocity spectrum of the recoil ions was measured by the time-of-flight method. The sliding gate B supported an alpha-particle source used for testing the two detectors.

region between the gate and the scintillator. The experimentally determined time-of-flight spectrum is represented by the histogram of Fig. 4. In contrast to the results of our first He^6 experiment, the new data appeared to agree with the tensor rather than the axial-vector curve.

The results of the experiment of Rustad and Ruby[4] became available before the completion of our second He^6 experiment. These authors measured the electron-neutrino angular correlation in the decay of He^6 and their results clearly indicated that the interaction was tensor. Since the statistical accuracy of this experiment was much higher than that obtainable with our method, we decided to give up our work on He^6 and to concentrate our efforts on other allowed decays.

EXPERIMENTS WITH A SINGLE ELECTROSTATIC SPECTROMETER

A recoil apparatus with improved energy resolution and transmission designed by Maxson and Allen[5] is shown in Fig. 5. The disintegration electrons were detected in a hemispherical plastic scintillator and a spherical electrostatic spectrometer was used to measure the energy of the recoils. The front surface of the source volume was defined, as in our previous experiments, by means of a sliding gate. This single spectrometer was used in our studies of Ne^{19} and A^{35}.

Ne^{19} is a positron emitter which decays according to the scheme

$$Ne^{19} \rightarrow F^{19} + e^+ + \nu, \quad \log ft = 3.27, \tag{5}$$

with a spin change of $J=\frac{1}{2}$ to $J=\frac{1}{2}$, no change in

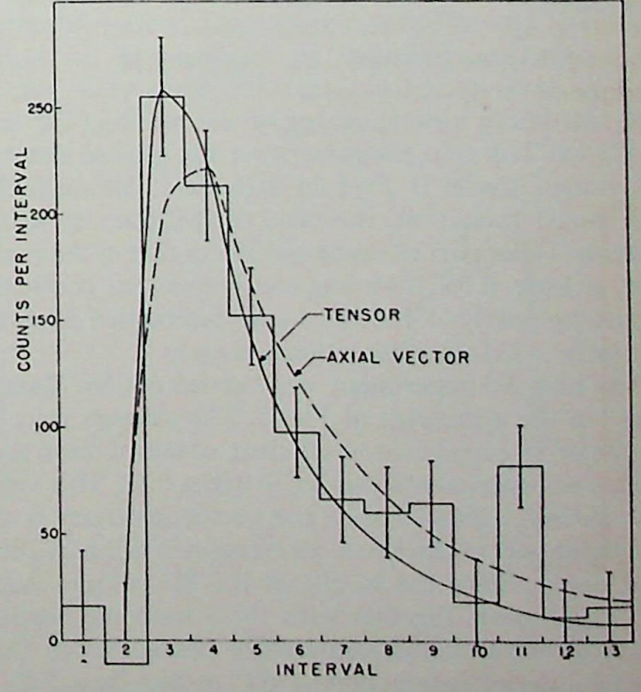


FIG. 4. The time-of-flight spectrum of the recoil ions from the decay of He^6. The angle between the directions of emission of the ions and the electrons varied from about 100° to 180°.

[4] B. M. Rustad and S. L. Ruby, Phys. Rev. **97**, 991 (1955).
[5] Maxson, Allen, and Jentschke, Phys. Rev. **97**, 109 (1955).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

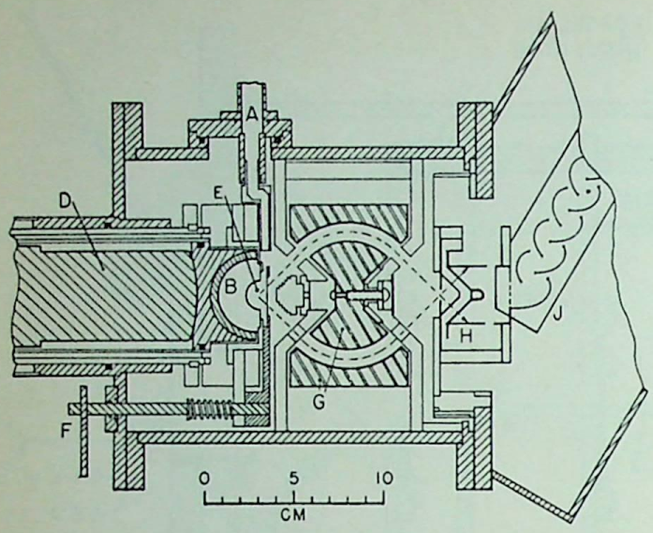


FIG. 5. The experimental arrangement used to study the decay of Ne^{19} and A^{35}. The effective source volume is at E. A spherical electrostatic spectrometer was used to measure the energy of the recoils.

nuclear parity. The small ft value indicates that this is an allowed transition. According to the selection rules for allowed transitions, this is a mixed decay with both Fermi and Gamow-Teller interactions present. The angular correlation coefficient computed from ft values is approximately -0.04 for the ST and $+0.04$ for the VA combination of interactions. The observed energy spectrum and several theoretical curves are shown in Fig. 6. A least squares fit of the theoretical curves to the experimental data yielded $\lambda=-0.21\pm0.08$, where the indicated error is statistical. This correlation was somewhat more negative than the value expected for the ST or VA interactions.

A^{35} is a positron emitter which decays according to the scheme

$$A^{35}\rightarrow Cl^{35}+e_{+}+\nu, \quad \log ft=3.79. \qquad (6)$$

The maximum kinetic energy of the recoiling Cl^{35} ions is 455 ev. The spin assignment for the ground states of the mirror nuclei is $J=\frac{3}{2}$ in each case. Although this is a mixed transition, the ratio of the intensity of the Gamow-Teller part of the transition to that of the Fermi part is only 0.05. The angular correlation coefficient should be nearly $+1$ if the Fermi interaction is vector or nearly -1 if the interaction is scalar.

The first A^{35} experiment was carried out by Maxson *et al.*[5] in the apparatus of Fig. 5. The average value of the angular correlation coefficient obtained from four groups of experiments was $\lambda=+0.85\pm0.12$. This value of λ clearly indicated that the vector invariant is the dominant part of the Fermi interaction in this particular beta decay. Since the results of the Rustad and Ruby He^{6} experiment, together with those from the neutron and Ne^{19} experiments, apparently demonstrated that the beta-decay interaction is ST rather than VT, a decision was made to repeat the A^{35} experiment with a new apparatus.

EXPERIMENTS WITH A DOUBLE SPECTROMETER

A second A^{35} experiment was performed by Herrmannsfeldt *et al.*[6] The energy distribution of the recoiling ions was measured directly without a simultaneous measurement of either the direction of motion or the energy of the associated electrons. The results were essentially the same as those from our first A^{35} experiment.

A schematic of the apparatus used in the second investigation of A^{35} is shown in Fig. 7. The two spectrometers were placed in series in order to provide space for differential pumping. The effective source volume for the first spectrometer was inside the cone at the left of the figure. The outer part of the source volume was defined by the metal wall of the cone and the inner part by the acceptance angle of the first spectrometer. Despite the use of differential pumping, the background counting rate was about ten times the maximum true rate. In order to measure this background, a set of grids was placed in front of the exit aperture of the source volume. A retarding potential could be placed on the second grid to shut off the beam of ions coming from the source volume. The net effect from the source was given by the difference between the counting rates observed with the retarding potential on and off. The intensity of the source was monitored by a scintillation counter viewing part of the source volume.

The experimental points and the theoretical recoil energy distributions for the decay of A^{35} are shown in Fig. 8. The experimental points correspond to an angular correlation factor of $\lambda=+0.97\pm0.14$ which is in excellent agreement with the value of $+0.98\pm0.08$ predicted for the AV interaction or the value of $+0.99\pm0.04$ expected for the TV interaction. In each case the predicted values of the angular correlation coefficient have been deduced from known ft values.

Since the results of our second A^{35} experiment con-

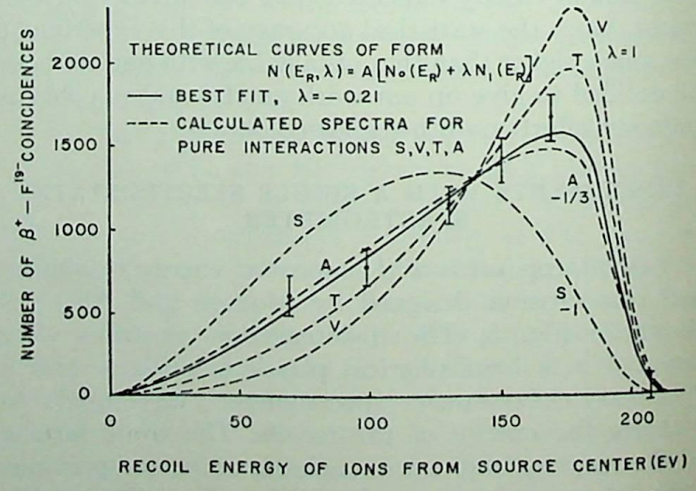


FIG. 6. The spectrum of the F^{19-} recoil nuclei from the Ne^{19} decay. The angular correlation parameter is -0.21 ± 0.08.

[6] Herrmannsfeldt, Maxon, Stahelin, and Allen, Phys. Rev. 107, 641 (1957).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

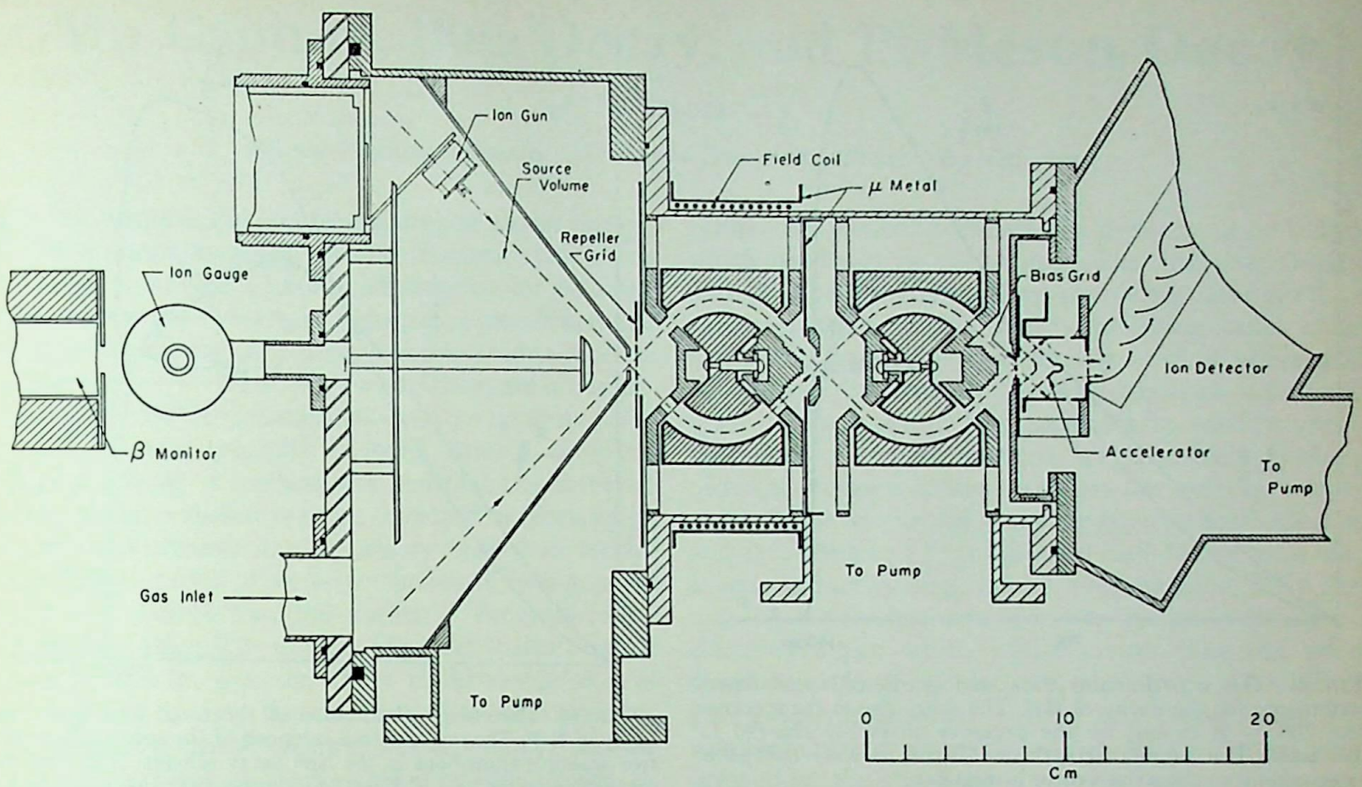


FIG. 7. A schematic of the double spectrometer used in the study of He^6, Ne^{19}, Ne^{23}, and A^{35}. The effective source volume is inside the cone at the left of the diagram. A retarding potential could be applied to a grid in front of the exit aperture of the source volume in order to separate the background from the true effect.

firmed those of the first, this was strong evidence for the VA interaction. Additional evidence for the axial-vector form appeared shortly after the completion of our second experiment. The most convincing evidence was supplied by Goldhaber *et al.*[7] who measured the helicity of the neutrino and found it to be left-handed. Their experiment clearly indicates that the Gamow-Teller interaction is axial-vector when it is supplemented by the information obtained from measurements of the longitudinal polarization of decay electrons. These developments suggested that a repetition of the He^6 experiment would be highly advisable.

A new He^6 experiment was carried out by Herrmannsfeldt *et al.*[8] using the double spectrometer of Fig. 7. The He^6 activity was produced through the $Be^9(n,\alpha)He^6$ reaction with fast neutrons inside the CP-5 reactor at the Argonne National Laboratory. The data from our latest He^6 experiment are shown in Fig. 9. The energy spectrum below one-half the maximum energy is distorted by the presence of doubly charged ions and cannot be used for a determination of λ. The angular correlation deduced from the data above 700 ev was $\lambda=-0.39\pm0.05$ corresponding to the axial-vector interaction.

This series of angular correlation experiments has been completed with a measurement of λ in the decay of Ne^{23}. This decay is complex with a 67% pure Gamow-Teller transition to the ground state of Na^{23} with $\Delta J=1$, no change of nuclear parity and a 32% transition to an excited state of Na^{23} with a probable spin change of $\Delta J=0$. Although this second transition is mixed, there are theoretical reasons for believing that the Fermi matrix element is very small. If this decay scheme is correct, the expected recoil energy spectrum should be a composite of two Gamow-Teller type spectra. Very weak transitions to higher levels of Na^{23}

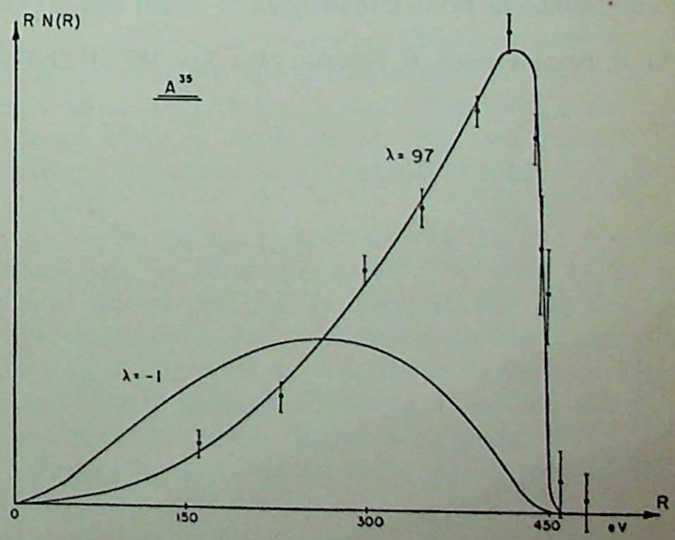


FIG. 8. The experimental data and predicted recoil energy distributions for the decay of A^{35}. The measured angular correlation factor is $\lambda=+0.97\pm0.04$ corresponding to the vector interaction.

[7] Goldhaber, Grodzins, and Sunyar, Phys. Rev. **109**, 1015 (1958).

[8] Herrmannsfeldt, Burman, Stahelin, Allen, and Braid, Phys. Rev. Letters **1**, 61 (1958).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

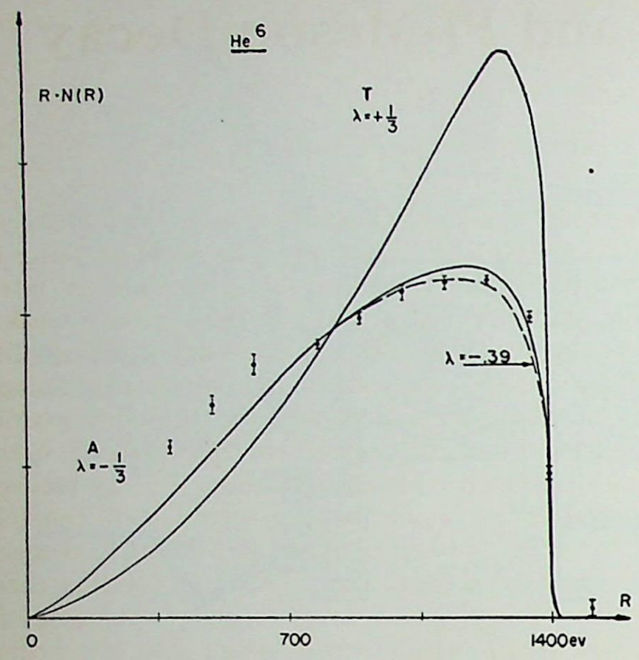


FIG. 9. The experimental data and predicted recoil energy distributions for the decay of He^6. The sharp rise in the spectrum below 700 ev is caused by the presence of doubly charged Li^6 recoil ions. The angular correlation factor is $\lambda=-0.39\pm0.05$ corresponding to the axial-vector interaction.

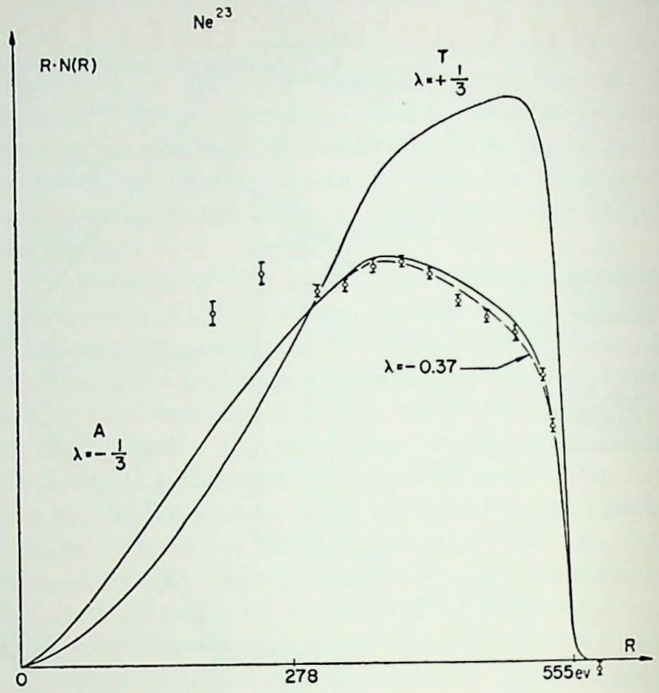


FIG. 10. The energy distribution of the recoil ions from the decay of Ne^{23}. The spectrum is a composit of the spectra from the two principle transitions in the Ne^{23} decay scheme. The angular correlation factor $\lambda=-0.37\pm0.04$ indicates that the axial-vector interaction is dominant in the Gamow-Teller type of decay.

can be neglected in the interpretation of this experiment. The measured spectrum is shown in Fig. 10. The angular correlation obtained from the upper one-half of the energy spectrum is $\lambda=-0.37\pm0.04$ where the indicated error is statistical.

The results of the Ne^{23} experiment can be interpreted in two different ways. If we assume that the $Ne^{23}-Na^{23}$ level scheme is correct, the experiment supplies additional evidence that the Gamow-Teller beta-decay interaction is axial-vector. The quoted uncertainty in the intensities of the two transitions[9] is sufficient to account for the observed deviation in λ from the value $-\frac{1}{3}$. Alternatively, we may consider the Gamow-Teller interaction to be predominantly axial-vector and in this case our experiment proves that the $Ne^{23}-Na^{23}$ level scheme is correct. Our results disagree with the value $\lambda=-0.05\pm0.10$ observed by Ridley.[10]

CONCLUSION

This series of experiments has removed most of the uncertainties concerning the nature of the beta-decay interaction which have been present for a number of years. We conclude that the dominant interaction form is *VA*. However, the present experiments do not exclude a possible tensor admixture of 10% or less. A scalar admixture of the same magnitude may also be present.

[9] J. R. Penning and F. H. Schmidt, Phys. Rev. **105**, 647 (1957).

[10] B. W. Ridley, Nuclear Phys. **6**, 34 (1958).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

Mu Capture, Beta Decay, and Pi-Meson Decay

M. L. GOLDBERGER

Palmer Physical Laboratory, Princeton University, Princeton, New Jersey

IN the original Yukawa formulation of meson theory, the π meson (as we now believe Yukawa's particle to be) was to provide a natural explanation for β decay. The process $\pi \rightarrow e+\bar{\nu}$ was regarded as an elementary interaction and nuclear β decay was imagined to proceed by the route $n \rightarrow p+\pi \rightarrow \mu+e+\bar{\nu}$. There are a variety of reasons why this scheme fails. Just the opposite point of view is now generally adopted, namely, that the nuclear β decay is fundamental and that the observed decay of the π meson is to be explained in terms of it. We do not exclude the possibility that β decay be described in terms of an as yet unknown heavy intermediate. Nevertheless, the nuclear β decay is to be regarded as essentially primary. In order to describe the actual dominant π-meson decay mode $\pi \rightarrow \mu+\bar{\nu}$ it is necessary to assume the existence of another β decay like process, μ-meson capture. The elementary process may be described as $\mu+p \rightarrow n+\nu$, or equally well as $n+\bar{p} \rightarrow \mu+\bar{\nu}$; the first is the experimentally observed μ-meson absorption reaction, whereas, the second, the annihilation of a neutron and an antiproton, plays an important role in π-meson decay.

Since this is a conference on weak interactions, I shall not be able to say that one of the π mesons, the neutral one, decays into two gamma rays; electromagnetic interactions are too strong to be mentioned! Furthermore, I will not be able to point out that a theory of the π^0 decay can be given which is very similar to what we describe for charged pions.

During the past year or two the field of weak interactions has become a surprisingly orderly one. The two-component theory of the neutrino, as well as the principle of lepton conservation now both seem to be well established. Both nuclear β decay and μ-meson decay seem to be describable in terms of a vector (V) and axial vector (A) coupling. This statement has to be qualified somewhat in the case of nuclear β decay. There is the additional fact that in both β decay and μ decay the vector coupling constants are almost identical. Rather less is known of the coupling types for the μ-meson capture reaction, but the dominant couplings seem to have about the same strength as in β decay. This is discussed by Primakoff. We tentatively assume that the apparently universal (V,A) interaction extends also to this Fermi process.

Precisely what do we mean by a *universal* interaction? This can mean only that the basic interaction Lagrangian contains only these two coupling types. Given this basic definition let us see whether there is anything surprising in the observed decays. First, in μ decay the V and A couplings are forced to be equal if we adopt the two component neutrino theory. In β decay, $g_A = 1.25\ g_V$, which need not be disturbing. The amazing thing is, with β decay and μ capture involving strongly interacting particles and μ decay involving only weakly interacting ones, that there is any kind of universality whatsoever. One would expect the existence of pions and other strongly interacting particles to modify greatly the effective matrix element for transitions between physical nucleons as compared to the μ-decay process.

Insofar as the vector coupling is concerned, Gerstein and Zel'dorich and Feynman and Gell-Mann have made a very attractive suggestion: They propose that there may be a principle analogous to gauge invariance in electrodynamics which would insure that the vector coupling constant in β decay be the same even when the strong interactions are turned on. Recall that as a result of current conservation, or, if you prefer, gauge invariance, the charge of a bare and physical proton is the same. In order to achieve this goal the β decay "vector current density" $g_V\bar{\psi}\gamma_\mu\psi$ [ψ is a nucleon field operator] must be augmented by terms which ultimately couple leptons directly to pions, etc., and such that the total "current density" $j_\mu{}^V$ satisfies $\partial j_\mu{}^V/\partial x_\mu$. The difference between vector and axial vector couplings in β decay is attributed to renormalization of the axial vector interaction.

One troublesome point in connection with this proposal has been raised by Wightman, Telegdi, and Michel. When one computes the electromagnetic radiative corrections for μ decay and β decay, one finds that to lowest order in all couplings not a finite correction for μ decay, but a logarithmic divergence in β decay. One may argue that if the nucleons are "dressed" properly and the radiative corrections are then computed (something no one knows how to do exactly) the result will be convergent. Nevertheless, it is not clear why, even if the β-decay effect is made finite, the two radiatively-corrected vector coupling constants should continue to be equal.

Let us discuss in a systematic way the role of strong interactions in Fermi processes. The work to be reviewed was carried out by Treiman and me and has been, for the most part, published elsewhere. I apologize for this, but in order to talk about something new, I would have to make an obviously wrong new theory—the correct one already having been given.

We suppose that β decay and μ capture are described by the Lagrangian density,

$$\mathcal{L}_I = Z_2 f_A \bar{\psi}_\nu(1-\gamma_5) i\gamma_\lambda\gamma_5\psi_1(\psi_n i\gamma_\lambda\gamma_5\psi_p) + Z_2 f_V \bar{\psi}_\nu(1-\gamma_5)\gamma_\lambda\psi_1(\bar{\psi}_n\gamma_\lambda\psi_p) + \text{Hermitian conjugate}, \quad (1)$$

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

where f_A and f_V are the unrenormalized coupling constants, and Z_2 is the nucleon wave-function renormalization constant. The ψ's are field operators associated with the particles indicated by the subscripts; l stands for either an electron or a μ meson. There may be other interactions of leptons. Among these are the direct pion couplings of Feynman and Gell-Mann, or perhaps couplings to baryons other than nucleons. For the time being we do not consider such possibilities. We consider the processes $(e,\mu)+p\rightarrow n+\nu$. To lowest order in the weak interaction, the matrix element computed from (1) is given by

$$S=i(2\pi)^4\delta(n+p_\nu-p-p_l)M, \tag{2}$$

where n, p_ν, p, p_l are the four-momenta of the neutron, neutrino proton, and electron (or μ meson), and

$$M=\bar{u}_\nu(1-\gamma_5)i\gamma_\lambda\gamma_5 u_l\langle n|P_\lambda|p\rangle + \bar{u}_\nu(1-\gamma_5)\gamma_\lambda u_l\langle n|V_\lambda|p\rangle \tag{3}$$

$|n\rangle$ and $|p\rangle$ represent physical neutron and proton states and

$$P_\lambda=Z_2 f_A\bar{\psi}_n i\gamma_\lambda\gamma_5\psi_p, \quad V_\lambda=Z_2 f_V\bar{\psi}_n\gamma_\lambda\psi_p. \tag{4}$$

The lepton spinors have been normalized according to $\bar{u}_l\gamma_4 u_l=\bar{u}_\nu\gamma_4 u_\nu=1$.

In the Feynman-Gell-Mann theory, V_λ would have additional terms. We do not use the explicit forms of V_λ and P_λ.

The general forms of the matrix elements of P_λ, V_λ required for Eq. (3) may be deduced from invariance principles, and they are

$$\langle n|P_\lambda|p\rangle=\left(\frac{m^2}{n_0p_0}\right)^{\frac{1}{2}}\bar{u}(n)\{ai\gamma_\lambda\gamma_5-b(p-n)_\lambda\gamma_5\}u(p), \tag{5}$$

$$\langle n|V_\lambda|p\rangle=\left(\frac{m^2}{n_0p_0}\right)^{\frac{1}{2}}\bar{u}(n)\{c\gamma_\lambda-d\sigma_{\lambda\mu}(p-n)_\mu\}u(p). \tag{6}$$

In these formulas, m is the nucleon mass, and the spinors are normalized according to $\bar{u}u=1$. The fact that only the momentum combination $p-n$ appears above is a consequence of charge symmetry and time reversal invariance in the strong interactions. Finally, the coefficients a, b, c, and d are functions of $(n-p)^2$, the momentum transfer squared.

Substituting these matrix elements in Eq. (3) and using the Dirac equation for the leptons we find for M the result

$$M=\left(\frac{m^2}{p_0n_0}\right)^{\frac{1}{2}}\{a\bar{u}_\nu(1-\gamma_5)i\gamma_\nu\gamma_5 u_l\bar{u}_n i\gamma_\lambda\gamma_5 u_p + m_1b\bar{u}_\nu(1-\gamma_5)\gamma_5 u_l\bar{u}_n\gamma_5 u_p+c\bar{u}_\nu(1-\gamma_5)\gamma_\lambda u_l\bar{u}_n\gamma_\lambda u_p + d\bar{u}_\nu(1-\gamma_5)\gamma_\lambda(p_l-p_\nu)_\mu u_l\bar{u}_n\sigma_{\lambda\mu}u_p\}.$$

The first and third terms are in the form of the usual axial vector and vector interactions. The functions $a[(n-p)^2]$ and $c[(n-p)^2]$ are, for zero value of the momentum transfer, simply the coupling constants g_A and g_V of β decay. In μ capture $(n-p)^2\sim m_\mu^2$ but a and c do not deviate much from g_A and g_V over such an interval. The second term has the form of a conventional pseudoscalar interaction with an effective coupling constant m_lb. Barring strong dependence of b on momentum transfer, this term is relatively much less important in β decay than in μ capture. The last term is identical with what has been called weak magnetism by Gell-Mann and is present whether or not the conserved vector current of Feynman and Gell-Mann is assumed. The magnitude of d depends critically, however, on this assumption.

This is as far as one can go on more or less general grounds. What we have done is to study the four functions a, b, c, and d by dispersion techniques. It is not practical to discuss this investigation in detail, so we outline the elements that go into such a treatment and quote the relevant results. We wish to represent the functions in the following form:

$$a(\xi)=g_A-\frac{\xi}{\pi}\int_{(3m_\pi)^2}^{\infty}d\xi'\frac{\mathrm{Im}a(-\xi')}{\xi'(\xi'+\xi-i\epsilon)},$$

$$b(\xi)=\frac{1}{\pi}\int_{m_\pi^2}^{\infty}d\xi'\frac{\mathrm{Im}b(-\xi')}{\xi'+\xi-i\epsilon},$$

$$c(\xi)=g-\frac{\xi}{\pi}\int_{(2m_\pi)^2}^{\infty}d\xi'\frac{\mathrm{Im}c(-\xi')}{\xi'(\xi'+\xi-i\epsilon)},$$

$$d(\xi)=\frac{1}{\pi}\int_{(2m_\pi)^2}^{\infty}d\xi'\frac{\mathrm{Im}d(-\xi')}{\xi'+\xi-i\epsilon}.$$

We have written the dispersion relations explicitly in such a way that $a(0)=g_A$, $c(0)=g_V$ and make essentially no effort to relate the renormalized coupling constants g_A, g_V to the unrenormalized ones appearing in the original Lagrangian. The quantities Ima, etc., represent the imaginary parts of the various amplitudes (which are real for positive arguments) and these may be expressed in terms of the amplitudes for certain real physical processes.

Consider first the vertex $\langle n|P_\lambda|p\rangle$; it is slightly more convenient to study $\langle 0|P_\lambda|\bar{n}p,in\rangle$ which is related to our other amplitude according to

$$\left(\frac{p_0\bar{n}_0}{m^2}\right)^{\frac{1}{2}}\langle 0|P_\lambda|n,\bar{p},in\rangle=\bar{v}_{\bar{n}}[ai\gamma_\lambda\gamma_5-b(p+\bar{n})_\lambda\gamma_5]u(p),$$

where $v_{\bar{n}}$ is a negative energy spinor and a is now a function of $(p+\bar{n})^2$. This matrix element may be imagined as describing the annihilation of a proton-antineutron pair to produce leptons via the interaction P_λ. The dependence on the lepton variables may be factored out so that they no longer appear explicitly. The sort of things that can contribute to this matrix element are shown in Fig. 1.

CC-0. Panjab University Chandigarh. Digitized by eGangotri University Haridwar Collection. Digitized by S3 Foundation USA

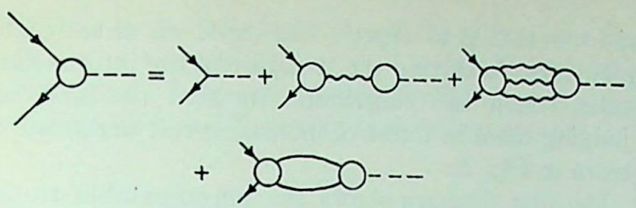


FIG. 1. The structure of the pseudoscalar matrix element $\langle 0|P_\lambda|n\bar{p}\rangle$ is pictured: The first "term" is effectively the bare interaction, the second shows a virtual transition to a π meson which decays into leptons, the third shows the reaction passing through a three pion state, and finally we have nucleon-antinucleon scattering followed by lepton emission.

The first element is essentially the bare interaction; in the second the pair annihilates to form a pion which then undergoes $\pi-\mu$ decay; the third diagram shows the pair annihilating into three pions, which ultimately combine to yield the lepton pair; the fourth diagram shows the pair undergoing a scattering interaction before annihilating to produce the leptons by the very matrix element we are studying. We are, thus, generating an integral equation. Needless to say, there are an infinite number of diagrams which we have not shown; we cannot even compute all of the ones we have. The intermediate state involving three pions is too hard for us to handle, but the remaining three are manageable and the integral equation for a and b can be easily solved.

The resulting solutions involve the ${}^3P_1$ and ${}^1S_0$ complex phase shifts for proton-antineutron scattering, the renormalized strong pion-nucleon coupling constant (G), the renormalized β-constant $[g_A=a(0)]$ and the experimental $\pi\rightarrow\mu+\nu$ lifetime. The latter enters via the one pion intermediate state which contributes only to the effective pseudoscalar interaction b. We find

$$a(\xi)=g_A\exp\left\{-\frac{\xi}{\pi}\int_{4m^2}^{\infty}dy\frac{\phi_1(y)}{y(y+\xi-i\epsilon)}\right\},$$

$$a-\frac{\xi}{2m}b=\left[g_A+\frac{\xi}{\xi+m_\pi{}^2}\frac{\sqrt{2}GF(-m_\pi{}^2)}{2m}\times\exp\left\{-\frac{m_\pi{}^2}{\pi}\int_{4m^2}^{\infty}dy\frac{\phi_0(y)}{y(y-m_\pi{}^2)}\right\}\right]\times\exp\left\{-\frac{\xi}{\pi}\int_{4m^2}^{\infty}dy\frac{\phi_0(y)}{y(y+\xi-i\epsilon)}\right\},$$

where $F(-m_\pi{}^2)$ may be related to the $\pi-\mu$ lifetime; ϕ_0 and ϕ_1 are related to the ${}^1S_0$ and ${}^3P_1$ phase shifts δ_0, δ_1 according to

$$\tan\phi=\frac{\mathrm{Re}e^{i\delta}\sin\delta}{1-\mathrm{Im}e^{i\delta}\sin\delta}.$$

We find, using the experimental value of the $\pi-\mu$ lifetime, $F=-0.115(\sqrt{2}Gmg_A/2\pi^2)$. Using this value and neglecting the contributions from the proton-antineutron scattering (i.e., set $\phi_0=\phi_1=0$) we find "g_p"$=m_\mu b(m_\mu{}^2)\approx 8g_A$ as the effective pseudoscalar coupling constant that would be effective in μ capture. The deviations of a from the value at $\xi=0$ are of order $m_\mu{}^2/m^2$; of course, the three pion intermediate state could cause slightly larger corrections but one would expect that for β decay or μ capture the leading terms

$$a\approx g_A,$$

$$b\approx-\frac{\sqrt{2}GF(-m_\pi{}^2)}{\xi+m_\pi{}^2},$$

are certainly adequate. In π decay, one needs a and b for values of $-\xi>4m^2$ in which case the neglect of the many less massive states (such as 3, $5\cdots$pions) could be much more serious. It is our feeling that since the leptons are coupled directly only to the nucleon pairs (or perhaps more generally to other baryon pairs) that such pair states are more important than the lighter pion states.

The effective vector interaction matrix element may be analyzed in a manner quite similar to our treatment of $\langle 0|P_\lambda|\bar{n}p,in\rangle$. We do not go into the analysis in much detail since the vector interaction plays no role in π decay. The matrix element $\langle 0|V_\lambda|\bar{n}p,in\rangle$ is identical in form to that encountered in the study of the electromagnetic structure of nucleons. If we follow Feynman and Gell-Mann we see that this parallel is essentially exact. As in the electromagnetic problem, the two-pion intermediate state is expected to play a dominant role for the relatively small values of $(\bar{n}+p)^2$, namely, about $m_\mu{}^2$ encountered in μ capture. In order to evaluate the two-pion contribution, we must know the matrix element for pair annihilation into two pions (even when the total energy extends into the unphysical region of total energy W, $4m^2>W^2>4\mu^2$) and also that for the pions to annihilate into a lepton pair. The latter process may also be analyzed by dispersion methods and we have done so in a rather crude fashion. For its evaluation one requires the matrix element for production of a proton-antineutron pair by two pions; the pair then annihilates via the original matrix element $\langle 0|V_\lambda|\bar{n}p,in\rangle$ (strictly speaking, there is a two pion intermediate state also which we neglect). We approximated all the matrix elements encountered in the problem by lowest order perturbation theory and found

$$c(\xi)=g_V\left[1+\frac{4}{9\pi}\frac{f^2}{4\pi}(0.24)\frac{\xi}{m_\pi{}^2}\right],$$

$$d(\xi)=1.7\frac{g_V}{2m}\times\frac{16}{3\pi}\frac{f^2}{4\pi}\left[1-\frac{0.12}{6}\frac{\xi}{m_\pi{}^2}\right],$$

where $f^2/4\pi=0.08$ is the effective pseudovector coupling constant of pion physics.

Adopting the Feynman—Gell-Mann theory, then, at least as far as the static terms (i.e., $\xi=0$) are concerned,

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

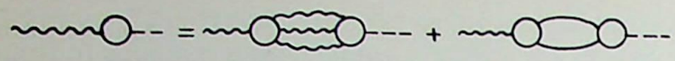


FIG. 2. Dispersion theoretic diagrams for π decay. The first shows a transition to an intermediate state with three pions, which we neglect; the second shows the usually contemplated transition through a nucleon-antinucleon pair.

the calculation can be carried out exactly. In their theory there is a complete analogy with the electromagnetic problem (except for slight numerical isotopic spin factors) so that we have without calculation,

$$c(0)\equiv g_V,$$
$$d(0)=(\mu_p-\mu_N)g_V/2m,$$

where μ_p, μ_N are the anomalous moments of proton and neutron in units of nucleon magnetons. Thus there is a clear-cut difference between the prediction of $d(0)$ made by the Feynman—Gell-Mann theory and the conventional theory: The value of d is about fifteen times larger in their case, and raises it out of the undetectable range.

An experiment to test the correctness of the Feynman—Gell-Mann proposal for conserved vector currents has been proposed by Gell-Mann. It is possible that the magnetic moment term, d, as well as the induced pseudoscalar interaction, b, may be detectable in certain μ-capture effects as discussed by Primakoff.

We turn now to a discussion of the decay of the π meson. It is some relief to be able to say that only the mode $\pi^-\to\mu^-+\bar{\nu}$ need be discussed. The correct branching ratio for the channel $\pi^-\to e^-+\bar{\nu}$ presumably then follows from our basic Lagrangian containing only vector and axial vector couplings. It is easy to see that the axial vector coupling alone plays any role in π decay. The leptons are assumed to emerge from a point (in the sense of an arbitrary Feynman diagram); hence there is only one momentum vector in the problem, say that of the pion, p_π. The pion is presumably a pseudoscalar and thus it is impossible to construct anything other than a pseudoscalar or a pseudovector to be coupled to the leptons. Formally, the S-matrix element for the transition is proportional to Λ where

$$\Lambda=\bar{u}(p_\mu)i\gamma_\lambda\gamma_5(1+\gamma_5)u(p_\nu)\langle 0|P_\lambda|\pi\rangle + \bar{u}(p_\mu)\gamma_\lambda(1+\gamma_5)u(p_\nu)\langle 0|V_\lambda|\pi\rangle$$

and the second term vanishes if one assumes parity is conserved in the strong interactions.

We concentrate attention, therefore, on $\langle 0|P_\lambda|\pi\rangle$ which we write as

$$\langle 0|P_\lambda|\pi\rangle=-i(p_\pi)_\lambda F(p_\pi{}^2)/(2p_{\pi 0})^{\frac{1}{2}}.$$

We can easily show that $F(p_\pi{}^2)$ satisfies a dispersion relation of the form

$$F(\xi)=\frac{1}{\pi}\int d\xi'\frac{\mathrm{Im}F(-\xi')}{\xi'+\xi-i\epsilon},$$

and our task is to express $\mathrm{Im}F$ in terms of calculable quantities. Analyzing the structure of the intermediate states which can contribute, we find the first few (judging them in terms of increasing rest mass) are as shown in Fig. 2.

The first diagram shows the uncomputable transition from one to three pions which ultimately combine to yield the leptons. There should then come states with 5, 7, $\cdots$ pions, perhaps followed by zero strangeness states involving K mesons and pions. Finally, one comes to the neutron-antiproton state. There are three reasons for concentrating attention on this state: (1) it is the one conventionally envisaged in a qualitative discussion of π-decay; (2) the leptons are directly coupled to nucleons, hence, such states might be expected to be of great importance; and (3) we can do quite a reasonable job of evaluating its contribution (and cannot calculate any of the others).

The individual pieces of our diagram are also treated by dispersion methods. We have already discussed the weak vertex in detail and so we now concentrate on the strong one, describing the virtual dissociation of the pion into a neutron-antiproton pair. In Fig. 3 the first diagram shows the "bare" interaction, the second, a three-pion state which by this time we neglect quite automatically, and, finally, the one we retain, namely, that involving a neutron-antiproton pair. This pair (in the rest system of the pion) is in a ${}^1S_0$ state (isotopic triplet) and to the indicated approximation can be characterized by a complex phase shift, δ_0.

For this vertex function, $K(\xi)$, say, one finds

$$K(\xi)=\sqrt{2}G\exp\left\{-\frac{\xi+m_\pi{}^2}{\pi}\int_{4m^2}^{\infty}d\xi'\frac{\phi_0(\xi')}{(\xi'-m_\pi{}^2)(\xi'+\xi-i\epsilon)}\right\},$$

where ϕ_0 is the same function introduced in connection with our previous discussion of $a-\xi b/2m$, the quantity arising from the weak vertex. Putting our dispersion pictures together, we find for $\mathrm{Im}F(\xi)$, neglecting small terms $\sim m_\pi{}^2/m^2$,

$$\mathrm{Im}F(\xi)=-\frac{\sqrt{2}G}{4\pi}\left[mg_A+\frac{\sqrt{2}GF(-m_\pi{}^2)\xi}{\xi+m_\pi{}^2}\right]\times\left[\frac{\xi+4m^2}{\xi}\right]^{\frac{1}{2}}H(\xi)$$
(for $-\xi>4m^2$, $=0$ otherwise),

where $H(\xi)$ is given by

$$H(\xi)=\exp\left\{-\frac{2(\xi+m_\pi{}^2)}{\pi}P\int_{4m^2}^{\infty}d\xi'\frac{\phi_0(\xi')}{(\xi'-m_\pi{}^2)(\xi'+\xi)}\right\},$$

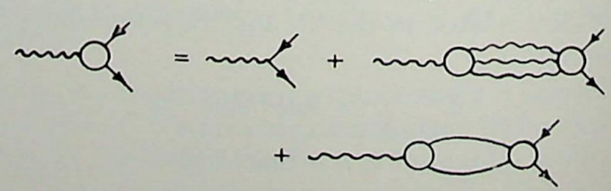


FIG. 3. Dispersion theoretic diagrams for the strong pion-nucleon vertex. The first one is the direct interaction, the second the uncomputable three pion state and finally the important state, the one involving a nucleon-antinucleon pair.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

and P means the principal value of the integral is to be taken at the singularity $\xi'=-\xi$. H is thus, aside from a factor, $|K(\xi)|^2$. On substituting $\mathrm{Im}F(\xi)$ into the dispersion relation for F, we find (neglecting m_π^2/m^2)

$$F(-m_\pi^2)\approx F(0)=-\frac{m\sqrt{2}}{4\pi}Gg_A\frac{J}{1+(G^2/4\pi)\cdot J},$$

where J is given by

$$J=\frac{1}{\pi}\int_{4m^2}^{\infty}d\xi\frac{(\xi-4m^2)^{\frac{1}{2}}}{\xi^3}H(-\xi).$$

If there were couplings to other baryon pairs (beside $\bar{p}n$) we would have

$$F(0)=-\frac{m}{4\pi}\sqrt{2}g_A\frac{\sum_i G_iJ_i}{1+(1/4\pi)\sum G_i^2J_i},$$

where the G_i are the strong coupling constants and the J_i are computed from H_i derived from the various ϕ_{0i}'s.

If we disregard the denominator and set the function H equal to unity, we get the familiar logarithmically-divergent result of perturbation theory. It has been conventional to say that if one puts the logarithm thus obtained equal to unity and computes the lifetime, the value of about sixty times the experimental one, which is found, is qualitative support for the correctness of the basic picture of π decay. The computation of the π-decay lifetime is really impossible for a person who believes seriously in the renormalization program based on perturbation theory. The π lifetime is a primitively divergent quantity whose presence must be accounted for by the existence of a so-called counter term which evidently then serves to remove the divergence and put in the observed decay rate by hand. One must prescribe the renormalized value of this divergent quantity. We obviously do not subscribe to this philosophy. Our feeling is that the function $H(\xi)$ plays a critical role and that the perturbation theoretical indications are irrelevant.

We obviously do not know enough about the complex $^1S_0$ phase shift for neutron-antiproton scattering to make a real quantitative study of J. What we have done, therefore, is to make a few simple models which have reasonable low energy behavior, and hope that they are not too insane at high energies. The reason that this may not be too unrealistic a procedure is that, provided only $H(\xi)\to 0$ however weakly for large ξ, the integral J exists. Furthermore, since it occurs in the denominator, multiplied by $G^2/4\pi\approx 15$, we see that if $J\gtrsim 1/15$ the J term dominates the denominator; neglecting the unity, then, J cancels out. This is a kind of strong coupling limit leading to $F(0)$ inversely proportional to G, instead of proportional to it, as would be given in weak coupling. In the case of several types of baryon loops there would evidently be a kind of mean value of $1/G$ defined by $\sum G_iJ_i/\sum G_i^2J_i$. In the global symmetry strong coupling limit ($G_i=G$) the earlier result continues to hold.

The models treated take for δ_0 the representation

$$\tan\delta_0=k(a+ib),\quad k=[(\xi/4)-m^2]^{\frac{1}{2}},$$

which leads to

$$\tan\phi_0=ka/(1+kb),$$

and we define $\delta_0(k=0)=0$. Various limiting cases of this expression have been studied ($a\gg b$, $b\gg a$) and in every case we find $J\gtrsim 0.7$. In a very unphysical case, namely, that of no absorption, $b=0$, all integrals may be evaluated analytically, and we find

$$J=-\frac{2}{\pi}\frac{ma+1}{ma-1}\{1-(m^2a^2-1)^{-\frac{1}{2}}\tan^{-1}(m^2a^2-1)^{\frac{1}{2}}\},$$

for the not unreasonable value of $ma\approx 3$ we find $J=0.7$.

Finally, making our strong coupling approximation, we obtain

$$F(0)=-(\sqrt{2}Gmg_A/\pi^2)[0.11],$$

using $G^2/4\pi=15$. This is to be compared with the experimental value given earlier, namely,

$$F(0)=-(\sqrt{2}Gmg_A/\pi^2)[0.115].$$

The agreement is rather impressive. It would be nice to hope that the neglect of all of the millions of states which we have made is contained in the 5% discrepancy. We are not quite so optimistic, but it is our feeling that the most important elements of the π-decay problem have been taken into account, and that a reasonable quantitative understanding of the process has been obtained.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

Theory of Muon Capture*

H. PRIMAKOFF

Department of Physics, Washington University, St. Louis 5, Missouri

THE present paper deals with various aspects of the theory of muon capture with emphasis on the relation between theory and experiment. The theory is based on an effective Hamiltonian, $H_{\text{eff}}^{(\mu)}$, which describes muon capture with subsequent neutrino emission by an aggregate of A dressed nucleons:

$$\mu^- + \left\{\begin{matrix} Z \text{ protons} \\ A-Z \text{ neutrons} \end{matrix}\right\} \rightarrow \nu + \left\{\begin{matrix} Z-1 \text{ protons} \\ A-Z+1 \text{ neutrons} \end{matrix}\right\}.$$

The treatment is subdivided as follows:

1. Effective Hamiltonian.
2. Total Muon Capture Rate in Closure Approximation—"Isotope" Effect.
3. Comparison of Closure Approximation Expression for Total Muon Capture Rate with Experiment.
4. Muon Capture to Particular Final States of Daughter Nucleus.
5. "Hyperfine" Effect in Muon Capture Rate—Muon Capture in Hydrogen.
6. Radiative Muon Capture: Total Rate and Photon-Neutrino Angular Correlation.
7. Parity Non-conservation Effects:

(a) Angular Distribution of Recoil Nuclei in Capture of Polarized Muons.
(b) Angular Distribution of Photons in Radiative Capture of Polarized Muons.
(c) Polarization of Recoil Nuclei in Muon Capture.
(d) Polarization of Photons in Radiative Muon Capture.

1. EFFECTIVE HAMILTONIAN

The appropriate expression for $H_{\text{eff}}^{(\mu)}$, as discussed in some detail by Fujii and Primakoff,[1] is, in a configuration space representation,

$$H_{\text{eff}}^{(\mu)} = \frac{1}{\sqrt{2}} \tau^{(+)} \frac{(1-\boldsymbol{\sigma}\cdot\boldsymbol{\nu}_1)}{\sqrt{2}} \sum_{i=1}^{A} \tau_i^{(-)} \{ G_V^{(\mu)} 1\cdot 1_i + G_A^{(\mu)} \boldsymbol{\sigma}\cdot\boldsymbol{\sigma}_i - G_P^{(\mu)} \boldsymbol{\sigma}\cdot\boldsymbol{\nu}_1 \boldsymbol{\sigma}_i\cdot\boldsymbol{\nu}_1 \} \delta(\mathbf{r}-\mathbf{r}_i) \quad (1a)$$

with

$$G_V^{(\mu)} \equiv g_V^{(\mu)} \left(1+\frac{\nu}{2m_p}\right);$$

$$G_A^{(\mu)} \equiv g_A^{(\mu)} - g_V^{(\mu)} (1+\mu_p-\mu_n) \frac{\nu}{2m_p}; \quad (1b)$$

$$G_P^{(\mu)} \equiv [g_P^{(\mu)} - g_A^{(\mu)} - g_V^{(\mu)} (1+\mu_p-\mu_n)] \frac{\nu}{2m_p}$$

and

$$g_V^{(\mu)} \cong g_V^{(\beta)} \times 0.972; \quad g_A^{(\mu)} \cong g_A^{(\beta)} \times 0.999; \quad g_P^{(\mu)} \cong 8 g_A^{(\mu)} \cong 8 g_A^{(\beta)}. \quad (1c)$$

This $H_{\text{eff}}^{(\mu)}$ corresponds,[1] in a nonrelativistic approximation for the muon and for the nucleons, to the most general Lorentz covariant transition matrix element for the reaction $\mu^- + p \rightarrow \nu + n$ in a theory where the lepton-bare nucleon coupling is V and A, where the neutrinos are emitted with unit negative helicity, where time reversal invariance holds, and where any bare hyperon, bare kayon "currents" which interact with the lepton current, have the same transformation property under the charge symmetry operation as the bare nucleon, bare pion currents.[2,3] However we neglect in Eq. (1) "many body" terms in $H_{\text{eff}}^{(\mu)}$ arising from the possibility of exchange of virtual pions, kayons, etc., among the nucleons. Such many body terms depend on the relative space coordinates of pairs, triplets, . . ., of nucleons and are believed, on the basis of a rough analysis of the corresponding beta decay situation, to be relatively small.[1]

In Eqs. (1a) to (1c), $g_V^{(\mu)}$, $g_A^{(\mu)}$, and $g_P^{(\mu)}$ are vector, axial vector and "induced" pseudoscalar muon-dressed nucleon coupling constants effective in muon capture while $g_V^{(\beta)}$ and $g_A^{(\beta)}$ are electron-dressed nucleon vector and axial vector coupling constants effective in beta decay. The numerical relations in Eq. (1c) between $g_V^{(\mu)}$, $g_V^{(\beta)}$; $g_A^{(\mu)}$, $g_A^{(\beta)}$ arise from the assumption of "universality" between the V, A muon-bare nucleon and electron-bare nucleon coupling constants which implies that $g_V^{(\mu)}$, $g_A^{(\mu)}$ differ from $g_V^{(\beta)}$, $g_A^{(\beta)}$ only because of the differing nucleon four-momentum transfers in the muon capture and in the beta decay. The numerical relation in Eq. (1c) between $g_P^{(\mu)}$ and $g_A^{(\mu)}$, due to Goldberger and Treiman[3] and to Wolfenstein,[4] is based on the assumption that a reaction such as: $\mu^- + \pi^+ \rightarrow \nu$ takes place predominantly via the sequence of "steps": $\mu^- + \pi^+ \rightarrow \mu^- + p + \bar{n} \rightarrow \nu$ which implies the possibility of muon capture via the "four-step process": $\mu^- + p \rightarrow \mu^- + \pi^+ + n \rightarrow \mu^- + p + \bar{n} + n \rightarrow \nu + n$. The quantities $\mu_p = 1.793$, $\mu_n = -1.913$ are the proton, neutron (static) anomalous magnetic moments (in units of $e/2m_p$ with e, m_p proton charge, mass and $\hbar = 1$, $c = 1$); these appear in the interaction effective in muon capture [Eq. (1b)] as a consequence of the Gell-Mann–

* Supported in part by the Air Force Office of Scientific Research of the Air Research and Development Command.

[1] A. Fujii and H. Primakoff, Nuovo cimento (to be published).

[2] S. Weinberg, Phys. Rev. 112, 1375 (1958).

[3] M. L. Goldberger and S. B. Treiman, Phys. Rev. 111, 355 (1958).

[4] L. Wolfenstein, Nuovo cimento 8, 882 (1958).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

Feynman assumption of a "conserved vector current" [5] which necessitates the existence, for example, of the reaction: $\mu^-+\pi^+\rightarrow\nu+\pi^0$ and hence implies the possibility of muon capture via the "three-step process": $\mu^-+p\rightarrow\mu^-+\pi^++n\rightarrow\nu+\pi^0+n\rightarrow\nu+n$. Also in Eqs. (1a) and (1b), $\mathbf{v}=\nu\mathbf{v}_1$ is the neutrino momentum; 1, 1_i and $\boldsymbol{\sigma}$, $\boldsymbol{\sigma}_i$ are 2×2 matrix unit operators and spin angular momentum operators for the lepton and the ith nucleon; $\mathbf{r}$ and $\mathbf{r}_i$ are space coordinates of the lepton and the ith nucleon; $\tau^{(+)}$, $\tau_i^{(-)}$ are isobaric-spin operators which transform a lepton muon state into a lepton neutrino state and an ith nucleon proton state into an ith nucleon neutron state; the factor $1/\sqrt{2}$ arises from the normalization of the neutrino relativistic wave function; the factor $(1-\boldsymbol{\sigma}\cdot\mathbf{v}_1)/\sqrt{2}$ is a consequence of the assumption of a maximum parity nonconserving two-component neutrino type muon-neutrino-nucleon coupling $(\sim(\psi_\nu{}^+[(1+\gamma_5)/\sqrt{2}]\gamma_4\gamma_\lambda\psi_\mu)\times(\psi_n{}^+\gamma_4\gamma_\lambda\psi_p)$, etc.). Though the muon and the nucleons are treated nonrelativistically in the derivation of $H_{\rm eff}{}^{(\mu)}$ all first-order nucleon recoil corrections, i.e., all terms in $H_{\rm eff}{}^{(\mu)}\sim\nu/m_p$, are nevertheless included.

2. TOTAL MUON CAPTURE RATE IN CLOSURE APPROXIMATION—"ISOTOPE" EFFECT

With the $H_{\rm eff}{}^{(\mu)}$ of Eqs. (1a) to (1c) we can obtain the square of the muon capture transition matrix element, $\langle|\text{M.E.}{}^{(\mu)}|^2\rangle$, summed over all spin orientations of the neutrino and averaged over all spin orientations of the muon. A straightforward calculation gives

$$\langle|\text{M.E.}{}^{(\mu)}|^2\rangle=\frac{1}{(2\pi)^3}\left[\frac{1}{\pi}\left(\frac{Zm_\mu{}'}{137}\right)^3\right]^{\frac{1}{2}}\cdot|\text{M.E.}_{\rm nucl}{}^{(\mu)}(a\rightarrow b)|^2, \quad (2a)$$

where $m_\mu{}'=m_\mu/[1+(m_\mu/Am_p)]$ is the muon reduced mass in the parent mu-mesic atom and the nuclear matrix element is expressed as

$$\begin{aligned}|\text{M.E.}_{\rm nucl}{}^{(\mu)}(a\rightarrow b)|^2&=(G_V{}^{(\mu)})^2|\langle b|\textstyle\sum_i\tau_i{}^{(-)}\exp(-i\nu_{ba}\cdot\mathbf{r}_i)\varphi(r_i)|a\rangle|^2\\&+(G_A{}^{(\mu)})^2|\langle b|\textstyle\sum_i\tau_i{}^{(-)}\exp(-i\nu_{ba}\cdot\mathbf{r}_i)\varphi(r_i)\boldsymbol{\sigma}_i|a\rangle|^2\\&+[(G_P{}^{(\mu)})^2-2G_A{}^{(\mu)}G_P{}^{(\mu)}]\\&\times|\langle b|\textstyle\sum_i\tau_i{}^{(-)}\exp(-i\nu_{ba}\cdot\mathbf{r}_i)\varphi(r_i)\boldsymbol{\sigma}_i\cdot\mathbf{v}_1|a\rangle|^2. \quad (2b)\end{aligned}$$

In Eqs. (2a) and (2b), $|a\rangle$, $|b\rangle$ represent wave functions of the two nuclear states involved in the capture process; the quantity $\varphi(r_i)$ is the muon space orbital wave function normalized in such a way that $\varphi(r_i)\rightarrow1$ as $Z\rightarrow0$, i.e., for small Z,

$$\varphi(r_i)\cong\exp(-Zm_\mu{}'r_i/137).$$

[5] R. P. Feynman and M. Gell-Mann, Phys. Rev. **109**, 193 (1958); M. Gell-Mann, Phys. Rev. **111**, 362 (1958); J. Bernstein and R. R. Lewis, Phys. Rev. **112**, 232 (1958).

The emitted neutrino momentum, ν_{ba}, is obtained from the energy and momentum conservation laws as

$$\nu_{ba}=m_\mu\left(1-\frac{\epsilon_a}{m_\mu}-\frac{(E_b-E_a)}{m_\mu}\right)\left(1-\frac{m_\mu}{2(m_\mu+Am_p)}\right), \quad (2c)$$

where ϵ_a is the binding energy of the muon in the lowest Bohr orbit of the mu-mesic atom and E_a, E_b are the energies of the nuclear states a, b. Thus, taking proper account of the density of final states available to the emitted neutrino, the total muon capture rate of the parent nucleus in the state $a|\rangle$, $\Lambda^{(\mu)}(a)$, is

$$\Lambda^{(\mu)}(a)=2\pi\cdot4\pi\sum_b\int\frac{d\mathbf{v}_1}{4\pi}\langle|\text{M.E.}{}^{(\mu)}|^2\rangle(\nu_{ba})^2\times[1+\nu_{ba}/[(\nu_{ba})^2+(Am_p)^2]^{\frac{1}{2}}]^{-1} \quad (3a)$$

so that using Eqs. (2a) to (2c)

$$\begin{aligned}\Lambda^{(\mu)}(a)=Z^3\frac{1}{2\pi^2}\frac{m_\mu{}^5}{(137)^3}\sum_b(\eta_{ba})^2\int\frac{d\mathbf{v}_1}{4\pi}&\times\{(G_V{}^{(\mu)})^2|\langle b|\textstyle\sum_i\tau_i{}^{(-)}\exp(-i\nu_{ba}\cdot\mathbf{r}_i)\varphi(r_i)|a\rangle|^2\\&+(G_A{}^{(\mu)})^2|\langle b|\textstyle\sum_i\tau_i{}^{(-)}\exp(-i\nu_{ba}\cdot\mathbf{r}_i)\varphi(r_i)\boldsymbol{\sigma}_i|a\rangle|^2\\&+[(G_P{}^{(\mu)})^2-2G_A{}^{(\mu)}G_P{}^{(\mu)}]\\&\times|\langle b|\textstyle\sum_i\tau_i{}^{(-)}\exp(-i\nu_{ba}\cdot\mathbf{r}_i)\varphi(r_i)\boldsymbol{\sigma}_i\cdot\mathbf{v}_1|a\rangle|^2\}\end{aligned} \quad (3b)$$

where

$$(\eta_{ba})^2\equiv\left(1-\frac{\epsilon_a}{m_\mu}-\frac{(E_b-E_a)}{m_\mu}\right)^2\left(1-\frac{m_\mu}{2(m_\mu+Am_p)}\right)^2\times[1+\nu_{ba}/[(\nu_{ba})^2+(Am_p)^2]^{\frac{1}{2}}]^{-1}\left(1+\frac{m_\mu}{Am_p}\right)^{-3}, \quad (3c)$$

or, dropping terms in m_μ/Am_p,

$$(\eta_{ba})^2\cong\left(1-\frac{\epsilon_a}{m_\mu}-\frac{(E_b-E_a)}{m_\mu}\right)^2\cong(\nu_{ba}/m_\mu)^2. \quad (3d)$$

The approximate expression for $(\eta_{ba})^2$ in Eq. (3d), involving neglect of the corrections for daughter nucleus recoil and parent mu-mesic atom reduced mass, holds to better than 5% for $A>12$.

The sum over b in Eqs. (3a) and (3b) runs over all energetically accessible states of the daughter nucleus, i.e., over all states b for which $E_b\leqq E_a+(m_\mu-\epsilon_a)$. Such energetically accessible states are very numerous since $m_\mu-\epsilon_\mu\approx100$ Mev and $(E_b)_{\text{ground state}}-E_a$ is normally ≈2–3 Mev and never exceeds 15 Mev; in addition, the matrix elements

$$\langle b|\textstyle\sum_i\tau_i{}^{(-)}\exp(-i\nu_{ba}\cdot\mathbf{r}_i)\varphi(r_i)|a\rangle,$$

$$\langle b|\textstyle\sum_i\tau_i{}^{(-)}\exp(-i\nu_{ba}\cdot\mathbf{r}_i)\varphi(r_i)\boldsymbol{\sigma}_i|a\rangle,$$

$$\langle b|\textstyle\sum_i\tau_i{}^{(-)}\exp(-i\nu_{ba}\cdot\mathbf{r}_i)\varphi(r_i)\boldsymbol{\sigma}_i\cdot\mathbf{v}_1|a\rangle$$

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

are largest for low lying states b. Thus one may have confidence in the accuracy of a closure approximation which extends the sum over all energetically accessible states b to a sum over all states b without restriction, and which replaces the explicitly E_b-dependent quantities, ν_{ba}, η_{ba}, by suitable averages $\langle\nu\rangle_a$, $\langle\eta\rangle_a$, which do not depend explicitly on E_b. Such a closure approximation applied to Eq. (3b) yields

$$\begin{aligned}\Lambda^{(\mu)}(a)=Z^3(\langle\eta\rangle_a)^2\frac{1}{2\pi^2}\frac{{m_\mu}^5}{(137)^3}\int\frac{d\boldsymbol{\nu}_1}{4\pi}\{&(G_V{}^{(\mu)})^2\langle a|\sum_{i,j}{}^{\cdot}\,\tau_i{}^{(+)}\tau_j{}^{(-)}\exp(i\langle\nu\rangle_a\boldsymbol{\nu}_1\cdot\mathbf{r}_{ij})\varphi^*(r_i)\varphi(r_j)|a\rangle\\&+(G_A{}^{(\mu)})^2\langle a|\sum_{i,j}\tau_i{}^{(+)}\tau_j{}^{(-)}\exp(i\langle\nu\rangle_a\boldsymbol{\nu}_1\cdot\mathbf{r}_{ij})\varphi^*(r_i)\varphi(r_j)\boldsymbol{\sigma}_i\cdot\boldsymbol{\sigma}_j|a\rangle\\&+((G_P{}^{(\mu)})^2-2G_P{}^{(\mu)}G_A{}^{(\mu)})\langle a|\sum_{i,j}\tau_i{}^{(+)}\tau_j{}^{(-)}\exp(i\langle\nu\rangle_a\boldsymbol{\nu}_1\cdot\mathbf{r}_{ij})\varphi^*(r_i)\varphi(r_j)\boldsymbol{\sigma}_i\cdot\boldsymbol{\nu}_1\boldsymbol{\sigma}_j\cdot\boldsymbol{\nu}_1|a\rangle\}\end{aligned}\tag{4a}$$

$$\begin{aligned}\cong Z^4(\langle\eta\rangle_a)^2\frac{1}{2\pi^2}\frac{{m_\mu}^5}{(137)^3}((G_V{}^{(\mu)})^2+3(\Gamma_A{}^{(\mu)})^2)\Bigg\{&\frac{\left\langle a\left|\sum_i\left(\frac{1+\tau_i{}^{(3)}}{2}\right)|\varphi(r_i)|^2\right|a\right\rangle}{\left\langle a\left|\sum_i\left(\frac{1+\tau_i{}^{(3)}}{2}\right)\right|a\right\rangle}\\&+\frac{\left\langle a\left|\sum_{i,j}{}'\left[\tfrac{1}{4}(\boldsymbol{\tau}_i\cdot\boldsymbol{\tau}_j-\tau_i{}^{(3)}\tau_j{}^{(3)})((G_V{}^{(\mu)})^2+(\Gamma_A{}^{(\mu)})^2\boldsymbol{\sigma}_i\cdot\boldsymbol{\sigma}_j)\right]\left[\frac{\sin(\langle\nu\rangle_a r_{ij})}{(\langle\nu\rangle_a r_{ij})}\varphi^*(r_i)\varphi(r_j)\right]\right|a\right\rangle}{Z((G_V{}^{(\mu)})^2+3(\Gamma_A{}^{(\mu)})^2)}\Bigg\}\end{aligned}$$

$$\begin{aligned}\equiv(Z_{\text{eff}})^4(\langle\eta\rangle_a)^2\frac{1}{2\pi^2}\frac{{m_\mu}^5}{(137)^3}&((G_V{}^{(\mu)})^2+3(\Gamma_A{}^{(\mu)})^2)\\\times\Bigg\{1+&\frac{\left\langle a\left|\sum_{i,j}{}'\left[\tfrac{1}{4}(\boldsymbol{\tau}_i\cdot\boldsymbol{\tau}_j-\tau_i{}^{(3)}\tau_j{}^{(3)})((G_V{}^{(\mu)})^2+(\Gamma_A{}^{(\mu)})^2\boldsymbol{\sigma}_i\cdot\boldsymbol{\sigma}_j)\right]\left[\frac{\sin(\langle\nu\rangle_a r_{ij})}{(\langle\nu\rangle_a r_{ij})}\varphi^*(r_i)\varphi(r_j)\right]\right|a\right\rangle}{Z((G_V{}^{(\mu)})^2+3(\Gamma_A{}^{(\mu)})^2)\left(\int|\varphi(r)|^2\mathfrak{D}_a(r)d\mathbf{r}\right)}\Bigg\},\end{aligned}$$

where

$$(\Gamma_A{}^{(\mu)})^2\equiv(G_A{}^{(\mu)})^2+\tfrac{1}{3}((G_P{}^{(\mu)})^2-2G_A{}^{(\mu)}G_P{}^{(\mu)});\quad\sum_{i,j}{}'\cdots\equiv\sum_{i,j;\,(i\neq j)}\cdots;\tag{4b}$$

$$(Z_{\text{eff}})^4\equiv Z^4\frac{\left\langle a\left|\sum_i\frac{1+\tau_i{}^{(3)}}{2}|\varphi(r_i)|^2\right|a\right\rangle}{\left\langle a\left|\sum_i\frac{1+\tau_i{}^{(3)}}{2}\right|a\right\rangle}=Z^4\int|\varphi(r)|^2\mathfrak{D}_a(r)d\mathbf{r}$$

with

$$\mathfrak{D}_a(r)\equiv\frac{1}{2J_a+1}\sum_{M_a}\frac{\left\langle E_a,J_a,M_a\left|\sum_i\left(\frac{1+\tau_i{}^{(3)}}{2}\right)\delta(\mathbf{r}-\mathbf{r}_i)\right|E_a,J_a,M_a\right\rangle}{\left\langle E_a,J_a,M_a\left|\sum_i\left(\frac{1+\tau_i{}^{(3)}}{2}\right)\right|E_a,J_a,M_a\right\rangle}.\tag{4c}$$

The approximate equality in Eq. (4a) refers to the replacement of

$$\int\frac{d\boldsymbol{\nu}_1}{4\pi}\left\langle a\left|\sum_{i,j}{}'\,\tau_i{}^{(+)}\tau_j{}^{(-)}\exp(i\langle\nu\rangle_a\boldsymbol{\nu}_1\mathbf{r}_{ij})\varphi^*(r_i)\varphi(r_j)\boldsymbol{\sigma}_i\cdot\boldsymbol{\nu}_1\boldsymbol{\sigma}_j\cdot\boldsymbol{\nu}_1\right|a\right\rangle\tag{4d}$$

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

by

$$\frac{1}{3}\int\frac{d\nu_1}{4\pi}\left\langle a\left|\sum_{i,j}{}'\,\tau_i^{(+)}\tau_j^{(-)}\frac{\sin(\langle\nu\rangle_a r_{ij})}{(\langle\nu\rangle_a r_{ij})}\varphi^*(r_i)\,\varphi(r_j)\boldsymbol{\sigma}_i\cdot\boldsymbol{\sigma}_j\right|a\right\rangle, \tag{4e}$$

a replacement which is justified since the d wave, g wave, $\cdots$ parts of $\exp(i\langle\nu\rangle_a\boldsymbol{\nu}_1\cdot\mathbf{r}_{ij})$ make a relatively small contribution to

$$\int\frac{d\nu_1}{4\pi}\left\langle a\left|\sum_{i,j}{}'\,\tau_i^{(+)}\tau_j^{(-)}\exp(i\langle\nu\rangle_a\boldsymbol{\nu}_1\cdot\mathbf{r}_{ij})\,\varphi^*(r_i)\,\varphi(r_j)\boldsymbol{\sigma}_i\cdot\boldsymbol{\nu}_1\boldsymbol{\sigma}_j\cdot\boldsymbol{\nu}_1\right|a\right\rangle.$$

The quantity $(Z_{\text{eff}})^4$, introduced by Wheeler in his original estimates of muon capture rates,[6] describes the variation of the muon space orbital wave function over the extent of the nucleus the quantity $\mathfrak{D}_a(r)$ being the (directionally-averaged-over) density function of the protons in the parent nucleus—the electrostatic potential appropriate to $\mathfrak{D}_a(r)$ enters into the muon energy eigenfunction-eigenvalue Dirac equation which determines ϵ_a and $\varphi(r)$. It is clear that $((Z_{\text{eff}})^4/Z^4)<1$, and $\to 1$ as $Z\to 0$.

Introducing the expression for the neutron decay rate,

$$\frac{\ln 2}{(\tau_{\frac{1}{2}})_{\text{neu}}}=\frac{1}{(2\pi)^3}m_e^5 f_{\text{neu}}\left[(g_V{}^{(\beta)})^2+3(g_A{}^{(\beta)})^2\right];\quad f_{\text{neu}}(\tau_{\frac{1}{2}})_{\text{neu}}=(1180\pm 35)\text{ sec},^7 \tag{5}$$

we have, from Eqs. (4a) to (4c) and (5),

$$\begin{aligned}\Lambda^{(\mu)}(a)&=(Z_{\text{eff}})^4(\langle\eta\rangle_a)^2\frac{(m_\mu/m_e)^5}{(137)^3}\frac{\pi(\ln 2)}{(f_{\text{neu}}(\tau_{\frac{1}{2}})_{\text{neu}})}\mathfrak{R}(1-\mathcal{g}_a)\\&=(Z_{\text{eff}})^4(\langle\eta\rangle_a)^2(272\text{ sec}^{-1})\mathfrak{R}(1-\mathcal{g}_a)\end{aligned} \tag{6a}$$

with

$$\mathfrak{R}\equiv\left[(G_V{}^{(\mu)})^2+3(\Gamma_A{}^{(\mu)})^2\right]/\left[(g_V{}^{(\beta)})^2+3(g_A{}^{(\beta)})^2\right], \tag{6b}$$

$$\mathcal{g}_a\equiv-\frac{\left\langle a\left|\sum_{i,j}{}'\left[\frac{1}{4}(\boldsymbol{\tau}_i\cdot\boldsymbol{\tau}_j-\tau_i^{(3)}\tau_j^{(3)})((G_V{}^{(\mu)})^2+(\Gamma_A{}^{(\mu)})^2\boldsymbol{\sigma}_i\cdot\boldsymbol{\sigma}_j)\right]\left[\frac{\sin(\langle\nu\rangle_a r_{ij})}{(\langle\nu\rangle_a r_{ij})}\varphi^*(r_i)\,\varphi(r_j)\right]\right|a\right\rangle}{Z((G_V{}^{(\mu)})^2+3(\Gamma_A{}^{(\mu)})^2)\left(\int|\varphi(r)|^2\mathfrak{D}_a(r)d\mathbf{r}\right)}. \tag{6c}$$

The quantity $\mathcal{g}_a(\mathcal{g}_a>0)$, which would vanish if the nucleus had $Z=A$ [in such a case: $(\boldsymbol{\tau}_i\cdot\boldsymbol{\tau}_j)|a\rangle=|a\rangle$; $\tau_i^{(3)}\tau_j^{(3)}|a\rangle=|a\rangle$] and also if $\langle\nu\rangle_a$ were $\gg\{(r_{ij})_{\text{Av}}\}^{-1}\approx$ mean nucleon momentum within nucleus, describes, within the context of the closure approximation, the inhibitory effect of the Pauli exclusion principle on the muon capture process. This inhibition may be visualized as arising from the fact that the neutron created in the $\mu^-+p\to n+\nu$ process cannot be produced in states already occupied by pre-existing neutrons of the parent nucleus, and the corresponding $\mathcal{g}_a$ may be expressed in terms of appropriate nucleon-nucleon correlation functions in the parent nucleus. We have

$$\begin{aligned}-\mathcal{g}_a=&\frac{\left\langle a\left|\sum_{i,j}{}'\left\{\frac{1}{4}(\boldsymbol{\tau}_i\cdot\boldsymbol{\tau}_j-\tau_i^{(3)}\tau_j^{(3)})\left[(G_V{}^{(\mu)})^2+(\Gamma_A{}^{(\mu)})^2\boldsymbol{\sigma}_i\cdot\boldsymbol{\sigma}_j\right]\right\}\left(\frac{1+P_{ij}}{2}\right)\right|a\right\rangle}{Z\left[(G_V{}^{(\mu)})^2+3(\Gamma_A{}^{(\mu)})^2\right]}\\&\times\int\int\frac{\sin(\langle\nu\rangle_a|\mathbf{r}-\mathbf{r}'|)}{(\langle\nu\rangle_a|\mathbf{r}-\mathbf{r}'|)}\varphi^*(r)\,\varphi(r')F_a{}^{(+)}(\mathbf{r},\mathbf{r}')d\mathbf{r}d\mathbf{r}'\Big/\int|\varphi(r)|^2\mathfrak{D}_a(r)d\mathbf{r}\\&+\frac{\left\langle a\left|\sum_{i,j}{}'\left\{\frac{1}{4}(\boldsymbol{\tau}_i\cdot\boldsymbol{\tau}_j-\tau_i^{(3)}\tau_j^{(3)})\left[(G_V{}^{(\mu)})^2+(\Gamma_A{}^{(\mu)})^2\boldsymbol{\sigma}_i\cdot\boldsymbol{\sigma}_j\right]\right\}\left(\frac{1-P_{ij}}{2}\right)\right|a\right\rangle}{Z\left[(G_V{}^{(\mu)})^2+3(\Gamma_A{}^{(\mu)})^2\right]}\\&\times\int\int\frac{\sin(\langle\nu\rangle_a|\mathbf{r}-\mathbf{r}'|)}{(\langle\nu\rangle_a|\mathbf{r}-\mathbf{r}'|)}\varphi^*(r)\,\varphi(r')F_a{}^{(-)}(\mathbf{r},\mathbf{r}')d\mathbf{r}d\mathbf{r}'\Big/\int|\varphi(r)|^2\mathfrak{D}_a(r)d\mathbf{r}\end{aligned} \tag{7a}$$

[6] J. A. Wheeler, Revs. Modern Phys. **21**, 133 (1949); J. Tiomno and J. A. Wheeler, Revs. Modern Phys. **21**, 153 (1949).
[7] Sosnovskii, Spivak, Prokofiev, Kutikov, and Dobrinin, J. Exptl. Theoret. Phys. (U.S.S.R.) **35**, 1059 (1958).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

with

$$F_a^{(\pm)}(\mathbf{r},\mathbf{r}')\equiv\frac{\left\langle a\left|\sum_{i,j}'\{\frac{1}{4}(\boldsymbol{\tau}_i\cdot\boldsymbol{\tau}_j-\tau_i^{(3)}\tau_j^{(3)})[(G_V^{(\mu)})^2+(\Gamma_A^{(\mu)})^2\boldsymbol{\sigma}_i\cdot\boldsymbol{\sigma}_j]\}[\delta(\mathbf{r}-\mathbf{r}_i)\delta(\mathbf{r}'-\mathbf{r}_j)]\left(\frac{1\pm P_{ij}}{2}\right)\right|a\right\rangle}{\left\langle a\left|\sum_{i,j}'\{\frac{1}{4}(\boldsymbol{\tau}_i\cdot\boldsymbol{\tau}_j-\tau_i^{(3)}\tau_j^{(3)})[(G_V^{(\mu)})^2+(\Gamma_A^{(\mu)})^2\boldsymbol{\sigma}_i\cdot\boldsymbol{\sigma}_j]\}\left(\frac{1\pm P_{ij}}{2}\right)\right|a\right\rangle},\tag{7b}$$

$$P_{ij}\equiv-\left(\frac{1+\boldsymbol{\sigma}_i\cdot\boldsymbol{\sigma}_j}{2}\right)\left(\frac{1+\boldsymbol{\tau}_i\cdot\boldsymbol{\tau}_j}{2}\right).\tag{7c}$$

The $F_a^{(\pm)}(\mathbf{r},\mathbf{r}')$ are nucleon-nucleon correlation density functions associated with a space symmetric, space antisymmetric relative motion of the two nucleons since for $|a\rangle$ antisymmetric in $\mathbf{r}_i$, $\mathbf{r}_j$; $\sigma_i^{(3)}$, $\sigma_j^{(3)}$; $\tau_i^{(3)}$, $\tau_j^{(3)}$; $P_{ij}|a\rangle=$ (exchange operator for $\mathbf{r}_i$, $\mathbf{r}_j$) $|a\rangle$; it is to be noted that $F_a^{(-)}(\mathbf{r},\mathbf{r})=0$. One has further

$$\begin{aligned}&\langle a|\sum_{i,j}'\{\tfrac{1}{4}(\boldsymbol{\tau}_i\cdot\boldsymbol{\tau}_j-\tau_i^{(3)}\tau_j^{(3)})[(G_V^{(\mu)})^2+(\Gamma_A^{(\mu)})^2\boldsymbol{\sigma}_i\cdot\boldsymbol{\sigma}_j]\}|a\rangle\\&\quad=-[(G_V^{(\mu)})^2+3(\Gamma_A^{(\mu)})^2]\tfrac{1}{2}A+\langle a|(G_V^{(\mu)})^2[(\mathbf{T})^2-(T^{(3)})^2]+(\Gamma_A^{(\mu)})^2[(\mathbf{Y}^{(1)})^2+(\mathbf{Y}^{(2)})^2]|a\rangle,\end{aligned}\tag{7d}$$

$$\begin{aligned}&\langle a|\sum_{i,j}'[\tfrac{1}{4}(\boldsymbol{\tau}_i\cdot\boldsymbol{\tau}_j-\tau_i^{(3)}\tau_j^{(3)})[(G_V^{(\mu)})^2+(\Gamma_A^{(\mu)})^2\boldsymbol{\sigma}_i\cdot\boldsymbol{\sigma}_j]P_{ij}|a\rangle\\&\quad=-[(G_V^{(\mu)})^2+3(\Gamma_A^{(\mu)})^2]\frac{Z(A-Z)}{2}+[(G_V^{(\mu)})^2-(\Gamma_A^{(\mu)})^2]\langle a|(\mathbf{S}_{\rm pr})^2+(\mathbf{S}_{\rm neu})^2-(\mathbf{S}_{\rm pr}+\mathbf{S}_{\rm neu})^2|a\rangle\end{aligned}$$

with

$$\mathbf{T}\equiv\sum_i\tfrac{1}{2}\boldsymbol{\tau}_i;\quad\mathbf{Y}^{(1),(2),(3)}\equiv\sum_i\tfrac{1}{2}\tau_i^{(1),(2),(3)}\boldsymbol{\sigma}_i;\quad\mathbf{S}_{\rm pr}\equiv\sum_i\left(\frac{1+\tau_i^{(3)}}{2}\right)(\boldsymbol{\sigma}_i/2);\quad\mathbf{S}_{\rm neu}\equiv\sum_i\left(\frac{1-\tau_i^{(3)}}{2}\right)(\boldsymbol{\sigma}_i/2)\tag{7e}$$

as may be verified by using relations of the type:

$$\sum_{i,j}'\tfrac{1}{4}\boldsymbol{\tau}_i\cdot\boldsymbol{\tau}_j=\sum_{i,j}\tfrac{1}{4}\boldsymbol{\tau}_i\cdot\boldsymbol{\tau}_j-\sum_i\tfrac{1}{4}(\boldsymbol{\tau}_i)^2=(\mathbf{T})^2-\tfrac{3}{4}A.$$

Equations (7a) to (7e) and (6a) to (6c) yield

$$\begin{aligned}\Lambda^{(\mu)}(a)=(Z_{\rm eff})^4(\langle\eta\rangle_a)^2(272\ \sec^{-1})\mathcal{R}\Bigg\{&1-\left[\frac{A}{2Z}-\frac{\langle a|(G_V^{(\mu)})^2[(\mathbf{T})^2-(T^{(3)})^2]+(\Gamma_A^{(\mu)})^2[(\mathbf{Y}^{(1)})^2+(\mathbf{Y}^{(2)})^2]|a\rangle}{Z[(G_V^{(\mu)})^2+3(\Gamma_A^{(\mu)})^2]}\right]\\&\times\left[\frac{\iint\frac{\sin(\langle\nu\rangle_a|\mathbf{r}-\mathbf{r}'|)}{(\langle\nu\rangle_a|\mathbf{r}-\mathbf{r}'|)}\varphi^*(r)\varphi(r')\tfrac{1}{2}[F_a^{(+)}(\mathbf{r},\mathbf{r}')+F_a^{(-)}(\mathbf{r},\mathbf{r}')]d\mathbf{r}d\mathbf{r}'}{\int|\varphi(r)|^2\mathfrak{D}_a(r)d\mathbf{r}}\right]\\&-\left[\frac{A-Z}{2}-\frac{[(G_V^{(\mu)})^2-(\Gamma_A^{(\mu)})^2]\langle a|(\mathbf{S}_{\rm pr})^2+(\mathbf{S}_{\rm neu})^2-(\mathbf{S}_{\rm pr}+\mathbf{S}_{\rm neu})^2|a\rangle}{Z[(G_V^{(\mu)})^2+3(\Gamma_A^{(\mu)})^2]}\right]\\&\times\left[\frac{\iint\frac{\sin(\langle\nu\rangle_a|\mathbf{r}-\mathbf{r}'|)}{(\langle\nu\rangle_a|\mathbf{r}-\mathbf{r}'|)}\varphi^*(r)\varphi(r')\tfrac{1}{2}[F_a^{(+)}(\mathbf{r},\mathbf{r}')-F_a^{(-)}(\mathbf{r},\mathbf{r}')]d\mathbf{r}d\mathbf{r}'}{\int|\varphi(r)|^2\mathfrak{D}_a(r)d\mathbf{r}}\right]\Bigg\}.\end{aligned}\tag{8}$$

Equation (8) for $\Lambda_{(a)}{}^{(\mu)}$ is essentially exact within the limitations of the closure approximation.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

We proceed to evaluate $\Lambda^{(\mu)}(a)$ for the heavier nuclei, $Z>6$, $A>12$. Dropping terms $\sim 1/Z$ in Eq. (8) yields

$$\Lambda^{(\mu)}(a)\cong(Z_{\text{eff}})^4(\langle\eta\rangle_a)^2(272\ \text{sec}^{-1})\mathfrak{R}\left\{1-\left(\frac{A}{2Z}\right)\left[\frac{\int\int\frac{\sin(\langle\nu\rangle_a|\mathbf{r}-\mathbf{r}'|)}{(\langle\nu\rangle_a|\mathbf{r}-\mathbf{r}'|)}\varphi^*(r)\varphi(r')\tfrac{1}{2}[F_a^{(+)}(\mathbf{r},\mathbf{r}')+F_a^{(-)}(\mathbf{r},\mathbf{r}')]d\mathbf{r}d\mathbf{r}'}{\int|\varphi(r)|^2\mathfrak{D}_a(r)d\mathbf{r}}\right]\right.$$
$$\left.-\left(\frac{A-Z}{2}\right)\left[\frac{\int\int\frac{\sin(\langle\nu\rangle_a|\mathbf{r}-\mathbf{r}'|)}{(\langle\nu\rangle_a|\mathbf{r}-\mathbf{r}'|)}\varphi^*(r)\varphi(r')\tfrac{1}{2}[F_a^{(+)}(\mathbf{r},\mathbf{r}')-F_a^{(-)}(\mathbf{r},\mathbf{r}')]d\mathbf{r}d\mathbf{r}'}{\int|\varphi(r)|^2\mathfrak{D}_a(r)d\mathbf{r}}\right]\right\}. \quad (9)$$

Then, adopting physically reasonable functional forms for $F_a^{(\pm)}(\mathbf{r},\mathbf{r}')$, $\mathfrak{D}_a(r)$ such as,

$$F_a^{(\pm)}(\mathbf{r},\mathbf{r}')=C_a^{(\pm)}\mathfrak{D}_a(r)\mathfrak{D}_a(r')(1\pm f_a(|\mathbf{r}-\mathbf{r}'|)); \quad f_a(|\mathbf{r}-\mathbf{r}'|)=\begin{matrix}1:|\mathbf{r}-\mathbf{r}'|\leqq d\\0:|\mathbf{r}-\mathbf{r}'|>d\end{matrix};$$
$$(C_a^{\pm})^{-1}=\int\int\mathfrak{D}_a(r)\mathfrak{D}_a(r')(1\pm f_a(|\mathbf{r}-\mathbf{r}'|))d\mathbf{r}d\mathbf{r}'; \quad \mathfrak{D}_a(r)=\begin{cases}1/[(4\pi/3)r_0^3A]:r\leqq r_0A^{\frac{1}{3}}\\0\qquad\qquad\qquad:r>r_0A^{\frac{1}{3}}\end{cases}; \quad (10)$$

remembering that $[\sin(\langle\nu\rangle_a|\mathbf{r}-\mathbf{r}'|)]/(\langle\nu\rangle_a|\mathbf{r}-\mathbf{r}'|)$ and $f_a(|\mathbf{r}-\mathbf{r}'|)$ are both *short ranged* functions with comparable ranges $(\langle\nu\rangle_a)^{-1}\approx d\approx r_0\ll r_0A^{\frac{1}{3}}$, and again neglecting terms $\sim 1/Z$, we have

$$\Lambda^{(\mu)}(a)\cong(Z_{\text{eff}})^4(\langle\eta\rangle_a)^2(272\ \text{sec}^{-1})\mathfrak{R}$$
$$\times\left\{1-\left(\frac{A-Z}{2A}\right)\left[\frac{A\int\int\frac{\sin(\langle\nu\rangle_a|\mathbf{r}-\mathbf{r}'|)}{(\langle\nu\rangle_a|\mathbf{r}-\mathbf{r}'|)}\varphi^*(r)\varphi(r')\mathfrak{D}_a(r)\mathfrak{D}_a(r')f_a(|\mathbf{r}-\mathbf{r}'|)d\mathbf{r}d\mathbf{r}'}{\int|\varphi(r)|^2\mathfrak{D}_a(r)d\mathbf{r}}\right]\right\}$$
$$\equiv(Z_{\text{eff}})^4(\langle\eta\rangle_a)^2(272\ \text{sec}^{-1})\mathfrak{R}\left\{1-\left(\frac{A-Z}{2A}\right)\delta_a\right\}. \quad (11a)$$

Further, using once more the *short ranged* character of $f_a(|\mathbf{r}-\mathbf{r}'|)$, we obtain

$$\delta_a\cong A\left(\frac{1}{(4\pi/3)r_0^3A}\right)\int|\varphi(r)|^2\mathfrak{D}_a(r)d\mathbf{r}\int\frac{\sin(\langle\nu\rangle_a\rho)}{(\langle\nu\rangle_a\rho)}f_a(\rho)d\rho\Big/\int|\varphi(r)|^2\mathfrak{D}_a(r)d\mathbf{r}$$
$$=\left(\frac{d}{r_0}\right)^3\left[1-\frac{1}{10}(\langle\nu\rangle_ar_0)^2\left(\frac{d}{r_0}\right)^2+\cdots\right]\cong\left(\frac{d}{r_0}\right)^3\left[1-0.024\left(\frac{d}{r_0}\right)^2+\cdots\right] \quad (11b)$$

where we have taken $\langle\nu\rangle_a\cong 0.75m_\mu$ (see below), $r_0\cong 1.25\times10^{-13}$ cm$=0.67/m_\mu$. Thus, within the present approximation, the "nucleon-nucleon correlation" parameter δ_a and the exclusion principle inhibition factor $\mathcal{I}_a$, proportional to the fraction of nucleons which are neutrons and to δ_a: $\mathcal{I}_a\cong((A-Z)/2A)\delta_a$ [Eqs. (6a), (6c), (11a), and (11b)], are both essentially determined by the parameter d/r_0, i.e., by the ratio of the characteristic lengths entering into $f_a(|\mathbf{r}-\mathbf{r}'|)$ and $\mathfrak{D}_a(r)$. The characteristic length d is the radius of the Pauli-Fermi "correlation region of influence," e.g., the Pauli-Fermi "correlation hole" surrounding each nucleon, and is determined by the interplay of the nucleon-nucleon forces and the exclusion principle. It is to be noted that up to terms $\sim 1/Z$, $\mathcal{I}_a$ vanishes with vanishing d, i.e., with vanishing nucleon-nucleon correlation.

To obtain the numerical value of d/r_0 and so of δ_a we consider the expression for the Coulomb energy of the parent nucleus. We have

$$E_{\text{Coul}}=\left\langle a\left|\sum_{i,j}{}'\frac{1}{2}\left(\frac{1+\tau_i^{(3)}}{2}\right)\left(\frac{1+\tau_j^{(3)}}{2}\right)\left(\frac{e^2}{r_{ij}}\right)\right|a\right\rangle, \quad (12)$$

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

and using a procedure analogous to that employed in the evaluation of $\mathcal{I}_a$ [Eqs. (6c) to (7e)] we obtain a formula first given by Feenberg and Goertzel,[8]

$$E_{\text{Coul}}=\tfrac{1}{2}Z(Z-1)\int\int\frac{e^2}{|\mathbf{r}-\mathbf{r}'|}\tfrac{1}{4}(G_a^{(+)}(\mathbf{r},\mathbf{r}')+3G_a^{(-)}(\mathbf{r},\mathbf{r}'))d\mathbf{r}d\mathbf{r}'$$
$$+(\tfrac{3}{4}Z-\langle a|(\mathbf{S}_{\text{pr}})^2|a\rangle)\int\int\frac{e^2}{|\mathbf{r}-\mathbf{r}'|}\tfrac{1}{2}(G_a^{(+)}(\mathbf{r},\mathbf{r}')-G_a^{(-)}(\mathbf{r},\mathbf{r}'))d\mathbf{r}d\mathbf{r}', \quad (13a)$$

where

$$G_a^{(\pm)}(\mathbf{r},\mathbf{r}')\equiv\frac{\left\langle a\left|\sum_{i,j}{}'\left[\left(\frac{1+\tau_i^{(3)}}{2}\right)\left(\frac{1+\tau_j^{(3)}}{2}\right)\right][\delta(\mathbf{r}-\mathbf{r}_i)\delta(\mathbf{r}'-\mathbf{r}_j)]\left(\frac{1+P_{ij}}{2}\right)\right|a\right\rangle}{\left\langle a\left|\sum_{i,j}{}'\left[\left(\frac{1+\tau_i^{(3)}}{2}\right)\left(\frac{1+\tau_j^{(3)}}{2}\right)\right]\left(\frac{1\pm P_{ij}}{2}\right)\right|a\right\rangle} \quad (13b)$$

is the proton-proton correlation density function associated with a space symmetric, space antisymmetric relative motion of the two protons. Comparing Eq. (13b) for $G_a^{\pm}(\mathbf{r},\mathbf{r}')$ with Eq. (7b) for $F_a^{(\pm)}(\mathbf{r},\mathbf{r}')$ we see that, in each case, the same spin and isobaric-spin operators occur in the integrals in the numerator and in the denominator so that the influence of these spin and isobaric-spin operators may be expected roughly to cancel. (This is almost rigorously true for nuclei in the $1s$ shell, e.g., He_2^4, where $|a\rangle$ factorizes into $\Phi_a(\mathbf{r}_1,\cdots,\mathbf{r}_4)\times X_a(\sigma_1^{(3)},\cdots,\sigma_4^{(3)};\tau_1^{(3)},\cdots,\tau_4^{(3)})$ to a good approximation.) It thus appears reasonable to identity $G_a^{(\pm)}(\mathbf{r},\mathbf{r}')$ with $F_a^{(\pm)}(\mathbf{r},\mathbf{r}')$ so that from Eqs. (13a) and (10),

$$E_{\text{Coul}}=(\tfrac{1}{2}Z(Z-1))\left(\frac{6}{5}\frac{e^2}{r_0A^{\frac{1}{3}}}\left(1-\frac{(d/r_0)^2}{A^{\frac{2}{3}}}\right)\right)+(\tfrac{3}{4}Z-\langle a|(\mathbf{S}_{\text{pr}})^2|a\rangle)\left(\frac{6}{5}\frac{e^2}{r_0A^{\frac{1}{3}}}\frac{(d/r_0)^2}{A^{\frac{2}{3}}}\right). \quad (13c)$$

A "best fit" of Eq. (13c) to the experimentally determined Coulomb energy differences of various light and medium-heavy mirror nuclei yields

$$\frac{d}{r_0}=1.47;\quad \delta_a=3.0\ [\text{from Eq. (11b)}]. \quad (14)$$

In this "best fit" the value of r_0 used is consistent with the electron-nucleus elastic scattering data[9]—the d/r_0 values of the individual nuclei naturally fluctuate somewhat about the aforementioned "best fit" value so that an assignment of, say, 10% uncertainty to the resultant value of 3.0 for δ_a appears quite reasonable. Equations (11a) and (14) yield values of the total muon capture rate, $\Lambda^{(\mu)}(a)$, which can be compared with experiment—it is to be noted that the effect of the exclusion principle inhibition is very important numerically since

$$\{1-\mathcal{I}_a\}\cong\left\{1-\left(\frac{A-Z}{2A}\right)\delta_a\right\}\cong\left\{1-\left(\frac{A-Z}{2A}\right)3\right\}=\frac{1}{4}$$

for Ca_{20}^{40}; $=\frac{1}{8}$ for Ca_{20}^{48}; $=5/32$ for Mo_{42}^{96}; $=19/208$ for Pb_{82}^{208}, etc. The associated *isotope effect* in the total muon capture rate is thus expected to be quite large, e.g., $\Lambda^{(\mu)}(Ca_{20}^{48})/\Lambda^{(\mu)}(Ca_{20}^{40})\cong\frac{1}{2}$, and would be interesting to verify experimentally using separated isotope targets.

Alternatively, one need not appeal to any relation between $F_a^{(\pm)}(\mathbf{r},\mathbf{r}')$ and $G_a^{(\pm)}(\mathbf{r},\mathbf{r}')$ with $G_a^{(\pm)}(\mathbf{r},\mathbf{r}')$ found from E_{Coul} in order to determine δ_a, but adopt instead the less ambitious procedure of fitting Eq. (11a) for $\Lambda^{(\mu)}(a)$ to the available data for the total muon capture rates, with use of a single Z, A-independent adjustable parameter δ_a. A successful fit of this type would also determine $(\langle\eta\rangle_a)^2\mathcal{R}$, i.e., with a reliable estimate of $\langle\eta\rangle_a\cong\langle\nu\rangle_a/m_\mu$ [Eq. (3d)] would determine $\mathcal{R}\equiv((G_V^{(\mu)})^2+3(\Gamma_A^{(\mu)})^2)/((g_V^{(\beta)})^2+3(g_A^{(\beta)})^2)$ [Eq. (6b)]. A more rigorous treatment would involve use of Eq. (8) for $\Lambda^{(\mu)}(a)$, without neglect of terms $\sim 1/Z$, with $\int|\varphi(r)|^2\mathcal{D}_a(r)d\mathbf{r}\equiv(Z_{\text{eff}})^4/Z^4$ appropriately expressed via a proton density function $\mathcal{D}_a(r)$ obtained from electron-nucleus elastic scattering data,[9] and with the nucleon-nucleon correlation density functions $F_a^{(\pm)}(\mathbf{r},\mathbf{r}')$, again identified with $G_a^{\pm}(\mathbf{r},\mathbf{r}')$ as deduced from an analysis of *inelastic* electron-nucleus scattering.[9] Such a more rigorous treatment must await the performance and the interpretation of the appropriate inelastic electron-nucleus scattering experiments.[10]

[8] E. Feenberg and G. Goertzel, Phys. Rev. **70**, 597 (1946).

[9] R. Hofstadter, Ann. Rev. Nuclear Sci. **7**, 231 (1957).

[10] H. A. Tolhoek, in a preprint received after this manuscript was completed, has applied the closure approximation to the calculation of $\Lambda^{(\mu)}(a)$ in a manner which, in effect, involves the expansion of $\exp(i\langle\nu\rangle_a\mathbf{v}_1\cdot\mathbf{r}_{ij})$ or of $\sin(\langle\nu\rangle_a r_{ij})/(\langle\nu\rangle_a r_{ij})$ [as in Eq. (4a) or (6c) or (8)] in powers of $\langle\nu\rangle_a r_{ij}\approx\langle\nu\rangle_a r_0A^{\frac{1}{3}}$. Tolhoek keeps terms only up to $(\langle\nu\rangle_a r_{ij})^2$ in this expansion. It is however to be emphasized that such an expansion converges rather slowly for the heavier nuclei ($Z>6$, $A>12$) and the results obtained by means of it are quite unreliable. It is to be noted that, contrary to the impression given by Tolhoek, no such expansion is used in the present procedure in the crucial passage from Eq. (8) to Eq. (11a).

CC-0. Kangri University Haridwar Collection. Digitized by S3 Foundation USA

TABLE I.

	$\langle\nu\rangle_a/m_\mu$	$(\langle\eta\rangle_a)^2$	$\langle a\vert(\mathbf{T})^2-(\mathbf{T}^{(3)})^2\vert a\rangle$	$\langle a\vert(\mathbf{Y}^{(1)})^2+(\mathbf{Y}^{(2)})^2\vert a\rangle$	$\langle a\vert(\mathbf{S}_{pr})^2+(\mathbf{S}_{neu})^2-(\mathbf{S}_{pr}+\mathbf{S}_{neu})^2\vert a\rangle$
H_1^1	0.94	0.58	1/2	3/2	0
H_1^2	0.90	0.65	0	2	−1/2
He_2^3	0.95	0.78	1/2	3/2	0
He_2^4	0.75	0.50	0	0	0

In concluding this section we apply Eq. (8) for $\Lambda^{(\mu)}(a)$ to the light nuclei H_1^1, H_1^2, He_2^3, He_2^4. We have from Table I and Eq. (8),

$$\Lambda^{(\mu)}(H_1^1)=1^4\times0.58\times272\ \text{sec}^{-1}\times\mathfrak{R}\times1,$$

$$\Lambda^{(\mu)}(H_1^2)\cong1^4\times0.65\times272\ \text{sec}^{-1}\times\mathfrak{R}\times\left\{1-\left[\frac{(G_V^{(\mu)})^2+(\Gamma_A^{(\mu)})^2}{(G_V^{(\mu)})^2+3(\Gamma_A^{(\mu)})^2}\right]\left[\int\int\frac{\sin(\langle\nu\rangle_a\vert\mathbf{r}-\mathbf{r}'\vert)}{(\langle\nu\rangle_a\vert\mathbf{r}-\mathbf{r}'\vert)}F_a^{(+)}(\mathbf{r},\mathbf{r}')d\mathbf{r}d\mathbf{r}'\right]\right\},\tag{15}$$

$$\Lambda^{(\mu)}(He_2^3)\cong2^4\times0.78\times272\ \text{sec}^{-1}\times\mathfrak{R}\times\left\{1-\frac{1}{2}\left[\int\int\frac{\sin(\langle\nu\rangle_a\vert\mathbf{r}-\mathbf{r}'\vert)}{(\langle\nu\rangle_a\vert\mathbf{r}-\mathbf{r}'\vert)}F_a^{(+)}(\mathbf{r},\mathbf{r}')d\mathbf{r}d\mathbf{r}'\right]\right\},$$

$$\Lambda^{(\mu)}(He_2^4)\cong2^4\times0.50\times272\ \text{sec}^{-1}\times\mathfrak{R}\times\left\{1-\left[\int\int\frac{\sin(\langle\nu\rangle_a\vert\mathbf{r}-\mathbf{r}'\vert)}{(\langle\nu\rangle_a\vert\mathbf{r}-\mathbf{r}'\vert)}F_a^{(+)}(\mathbf{r},\mathbf{r}')d\mathbf{r}d\mathbf{r}'\right]\right\},$$

which exhibits forms for the corresponding exclusion principle inhibition factors, $\mathcal{I}_a$ [Eqs. (15) and (6a) to (6c)]. With regard to the first two columns of Table I for $\langle\nu\rangle_a$ [Eq. (2c)] and $\langle\eta\rangle_a$ [Eq. (3c)] reasonable estimates have been made for $\langle E_b-E_a\rangle=\langle$excitation energy of daughter nucleus $b\rangle+((E_b)_{\text{ground state}}-E_a)$; also $\varphi(r)\approx\exp(-Zm_\mu' r/137)$ has been approximated by 1 within the integrals. Further, from Eq. (7b), the $\mathbf{r}_i$ being here coordinates relative to the nucleus' center of mass,

$$F_a^{(+)}(\mathbf{r},\mathbf{r}')=\int\int|\Phi_a(\mathbf{r}_1,\mathbf{r}_2)|^2\delta(\mathbf{r}_1+\mathbf{r}_2)\delta(\mathbf{r}-\mathbf{r}_1)\delta(\mathbf{r}'-\mathbf{r}_2)d\mathbf{r}_1d\mathbf{r}_2=|\Phi_a(\mathbf{r},\mathbf{r}')|^2\delta(\mathbf{r}+\mathbf{r}'):H_1^2$$

$$F_a^{(+)}(\mathbf{r},\mathbf{r}')=\int\int\int|\Phi_a(\mathbf{r}_1,\mathbf{r}_2,\mathbf{r}_3)|^2\delta(\mathbf{r}_1+\mathbf{r}_2+\mathbf{r}_3)\delta(\mathbf{r}-\mathbf{r}_1)\delta(\mathbf{r}'-\mathbf{r}_2)d\mathbf{r}_1d\mathbf{r}_2d\mathbf{r}_3=|\Phi_a[\mathbf{r},\mathbf{r}',-(\mathbf{r}+\mathbf{r}')]|^2:He_2^3\tag{16}$$

$$F_a^{(+)}(\mathbf{r},\mathbf{r}')=\int\int\int\int|\Phi_a(\mathbf{r}_1,\mathbf{r}_2,\mathbf{r}_3,\mathbf{r}_4)|^2\delta(\mathbf{r}_1+\mathbf{r}_2+\mathbf{r}_3+\mathbf{r}_4)\delta(\mathbf{r}-\mathbf{r}_1)\delta(\mathbf{r}'-\mathbf{r}_2)d\mathbf{r}_1d\mathbf{r}_2d\mathbf{r}_3d\mathbf{r}_4$$

$$=\int|\Phi_a[\mathbf{r},\mathbf{r}',\mathbf{r}_3,-(\mathbf{r}+\mathbf{r}'+\mathbf{r}_3)]|^2d\mathbf{r}_3:He_2^4,$$

where the Φ_a are space wave functions in the corresponding H_1^2, He_2^3, He_2^4 space, spin, isobaric-spin wave functions $|a\rangle\cong\Phi_a(\mathbf{r}_1,\mathbf{r}_2,\cdots)X_a(\sigma_1^{(3)},\sigma_2^{(3)},\cdots;\tau_1^{(3)},\tau_2^{(3)},\cdots)$ with $\Phi_a(\mathbf{r}_1,\mathbf{r}_2,\cdots)=\Phi_a(r_{12},\cdots)=P_{ij}\Phi_a(\mathbf{r}_1,\mathbf{r}_2,\cdots)$. The appropriate spin, isobaric-spin wave functions, X_a, have been used together with Eq. (7e) in order to obtain the entries in the last three columns of Table I. Use of simple variational trial forms for Φ_a enables evaluation of the $F_a^{(+)}(\mathbf{r},\mathbf{r}')$ in Eq. (16) which, together with the $\langle\nu\rangle_a$ estimates of the first column, give the integrals in Eq. (15). Employing also numerical values of $\mathfrak{R}$ quoted below [Eq. (19)] and a value of $(\Gamma_A^{(\mu)}/G_V^{(\mu)})^2=1.53$ calculated from Eqs. (1b), (1c), and (4b) with $(g_A^{(\beta)}/g_V^{(\beta)})=-1.21$,[11] we obtain

$$\Lambda^{(\mu)}(H_1^1)=1^4\times0.58\times272\ \text{sec}^{-1}\times\mathfrak{R}\times1=158\ \text{sec}^{-1}\times\mathfrak{R}=169\ \text{sec}^{-1},$$

$$\Lambda^{(\mu)}(H_1^2)\cong1^4\times0.65\times272\ \text{sec}^{-1}\times\mathfrak{R}\times\left\{1-\left[\frac{(G_V^{(\mu)})^2+(\Gamma_A^{(\mu)})^2}{(G_V^{(\mu)})^2+3(\Gamma_A^{(\mu)})^2}\right]\times0.64\right\}$$

$$=177\ \text{sec}^{-1}\times\mathfrak{R}\times\left\{1-\left[\frac{1+(\Gamma_A^{(\mu)}/G_V^{(\mu)})^2}{1+3(\Gamma_A^{(\mu)}/G_V^{(\mu)})^2}\right]\times0.64\right\}=135\ \text{sec}^{-1}\tag{17}$$

$$\Lambda^{(\mu)}(He_2^{(3)})\cong2^4\times0.78\times272\ \text{sec}^{-1}\times\mathfrak{R}\times(1-\tfrac{1}{2}\times0.66)=23\times10^2\ \text{sec}^{-1}\times\mathfrak{R}=25\times10^2\ \text{sec}^{-1}$$

$$\Lambda^{(\mu)}(He_2^{(4)})\cong2^4\times0.50\times272\ \text{sec}^{-1}\times\mathfrak{R}\times(1-0.80)=4.4\times10^2\ \text{sec}^{-1}\times\mathfrak{R}=4.7\times10^2\ \text{sec}^{-1}.$$

[11] C. S. Wu, Revs. Modern Phys. 30, 783 (1959), this issue; V. L. Telegdi, Conference on Weak Interactions, Gatlinburg, Tennessee (1958).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

The dependence of $\Lambda^{(\mu)}(H_1^2)$ on the ratio $(\Gamma_A{}^{(\mu)}/G_V{}^{(\mu)})^2$ is especially to be noted[12] as is the enormous isotope effect—factor $\cong 5$—between $\Lambda^{(\mu)}(He_2^3)$ and $\Lambda^{(\mu)}(He_2^4)$. Both of these are essentially manifestations of the inhibitions of the exclusion principle on the total muon capture rate. Thus in $\mu^-+H_1^2 \rightarrow \nu+n+n$ the dineutron must be produced in the ${}^3P_1$ state if $\Gamma_A{}^{(\mu)}=0$ while for $|\Gamma_A{}^{(\mu)}| \gtrsim G_V{}^{(\mu)}$ production in the dineutron ${}^1S_0$ state is possible. Since the dineutron ${}^1S_0$ state spatially overlaps far better with the deuteron ground ${}^3S_1$ state than does the dineutron ${}^3P_1$ state, dineutron production in the ${}^1S_0$ state when $|\Gamma_A{}^{(\mu)}| \gtrsim G_V{}^{(\mu)}$ is indeed predominant, and one expects $\Lambda^{(\mu)}(H_1^2)$ to be greater when $(\Gamma_A{}^{(\mu)}/G_V{}^{(\mu)})^2 \gg 1$ than when $(\Gamma_A{}^{(\mu)}/G_V{}^{(\mu)})^2 \ll 1$ just as predicted by Eq. (17). In a similar way the factor of 5 difference between the rates of $\mu^-+He_2^3 \rightarrow \nu+H_1^3$ and $\mu^-+He_2^4 \rightarrow \nu+H_1^4$ may be viewed as largely arising from the fact that the H_1^3 may be formed in ${}^2S_{\frac{1}{2}}$ states (bound or unbound) which spatially overlap well with the ${}^2S_{\frac{1}{2}}$ ground state of He_2^3 while the H_1^4 is formed in (unbound) ${}^3P_{2,1,0}$, ${}^1P_0$ states which all have a poor spatial overlap with the ${}^1S_0$ ground state of He_2^4.

3. COMPARISON OF CLOSURE APPROXIMATION EXPRESSION FOR THE TOTAL MUON CAPTURE RATE WITH EXPERIMENT

Equation (11a) for $\Lambda^{(\mu)}(a)$ has been compared by Telegdi, Sens, Swanson, and Yovanovitch[13] with their definitive measurements of the total muon capture rates in 29 elements from C_6^{12} to U_{92}^{238}. These investigators first calculated values of $(Z_{\text{eff}})^4$ [according to Eq. (4c)] for the various elements studied and then found that a plot of their $[\Lambda^{(\mu)}(a)]_{\text{exper}}/(Z_{\text{eff}})^4$ values *versus* the corresponding $(A-Z)/2A$ values gave a nice straight line in agreement with Eq. (11a); the values of δ_a and $[(\langle\eta_a\rangle_a)^2(272\ \text{sec}^{-1})\mathfrak{R}]$, determined by a weighted least squares fit of their individual experimental points to this straight line, were

$$\delta_a=3.15 \quad [vs\ \delta_a=3.0 \text{ in Eq. (14)}] \qquad (18a)$$

and

$$(\langle\eta\rangle_a)^2(272\ \text{sec}^{-1})\mathfrak{R}=188\ \text{sec}^{-1}\ \text{(experimental)}. \qquad (18b)$$

On the other hand, Eqs. (1b), (1c), (4b), (6b), and (3d), and $(g_A{}^{(\beta)}/g_V{}^{(\beta)})=-1.21$ [11] yield:

$$\mathfrak{R}=\begin{matrix}1.06:\langle\eta\rangle_a\cong\langle\nu\rangle_a/m_\mu\cong 0.75\\ 1.07:\langle\eta\rangle_a\cong\langle\nu\rangle_a/m_\mu\cong 0.85-0.95,\end{matrix} \qquad (19)$$

while the best *a priori* estimate for

$$\langle\eta\rangle_a\cong\langle\nu\rangle_a/m_\mu\cong\{1-(\epsilon_a/m_\mu -\langle\text{excitation energy of daughter nucleus } b\rangle/m_\mu -[(E_b)_{\text{ground state}}-E_a]/m_\mu\},$$

[Eq. (3d)] supposed valid in the mean for all the various pairs of nuclei (b,a) involved, is $\langle\eta\rangle_a\cong\langle\nu\rangle_a/m_\mu\cong 0.75=80$ Mev/m_μ with an uncertainty of, say, 10%. This value of $\langle\nu\rangle_a/m_\mu$ corresponds to an $\langle$excitation energy of daughter nucleus $b\rangle\cong 15$ Mev which is of the order of those empirically observed. The value of $(\langle\eta\rangle_a)^2(272\ \text{sec}^{-1})\mathfrak{R}$ for $\langle\eta\rangle_a\cong\langle\nu\rangle_a/m_\mu\cong 0.75=80$ Mev/m_μ is, using Eq. (19);

$$(\langle\eta\rangle_a)^2(272\ \text{sec}^{-1})\mathfrak{R}=161\ \text{sec}^{-1}\ \text{(theoretical)} \qquad (20)$$

which, in view of the uncertainty in the value of $(\langle\eta\rangle_a)^2$, must be considered in essential agreement with the experimental value of 188 sec^{-1} [Eq. (18b)]—thus, for example, the not unreasonable choice of $\langle\nu\rangle_a/m_\mu \cong 0.80=85$ Mev/m_μ yields $(\langle\eta\rangle_a)^2(272\ \text{sec}^{-1})\mathfrak{R}=185$ sec^{-1} (theoretical). In this way one finds support for the combination of basic assumptions on which our effective Hamiltonian $H_{\text{eff}}{}^{(\mu)}$ [Eqs. (1a) to (1c)] rests, *viz.*: (a) "universality" between muon-bare nucleon and electron-bare nucleon coupling constants which implies the numerical relations of Eq. (1c) between the effective muon-dressed nucleon and electron-dressed nucleon coupling constants; (b) the presence of an "induced" pseudoscalar interaction[3,4] with an effective muon-dressed nucleon coupling constant $g_P{}^{(\mu)}=8g_A{}^{(\beta)}$ [Eq. (1c)]; and (c) the presence of anomalous nucleon magnetic moment contributions in the effective muon-dressed nucleon interaction associated with the assumption of a "conserved vector current." [5] In particular, if this assumption of the conserved vector current is abandoned and the anomalous nucleon magnetic moment contributions $\sim(\mu_p-\mu_n)$ to the effective muon-dressed coupling constants $G_A{}^{(\mu)}$, $G_p{}^{(\mu)}$ omitted from Eqs. (1b), (4b), and (6b), the quantity $\mathfrak{R}$ of Eq. (6b) is 0.90 rather than 1.06 and

$$(\langle\eta\rangle_a)^2(272\ \text{sec}^{-1})\mathfrak{R}=137\ \text{sec}^{-1}\ \text{(theoretical)}. \qquad (20)$$

It thus appears that somewhat better agreement is reached between theoretical and experimental values of the total muon capture rates if the assumption of a conserved vector current is retained.

4. MUON CAPTURE TO PARTICULAR FINAL STATES OF DAUGHTER NUCLEUS

The first investigation in which a *partial* muon capture rate was determined is due to Godfrey.[14] Godfrey studied experimentally the rate of that muon capture reaction

$$\mu^-+C_6^{12} \rightarrow \nu+B_5^{12}$$

[12] H. Primakoff, Phys. Rev. **91**, 480 (A) (1953), and *Proceedings of the Fifth Annual Rochester Conference on High Energy Physics*, New York (1955), p. 174; A. Rudik, Doklady Akad. Nauk. **92**, 739 (1953); H. Überall and L. Wolfenstein, Nuovo cimento **10**, 136 (1958).

[13] V. Telegdi (private communication); Sens, Swanson, Telegdi, and Yovanovitch, Phys. Rev. **107**, 1464 (1957); J. Sens, Ph.D. thesis, University of Chicago (1958) and Phys. Rev. **113**, 679 (1959).

[14] T. N. K. Godfrey, Ph.D. thesis, Princeton University (1954), and Phys. Rev. **92**, 512 (1953).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

which was followed by the beta decay of B_5^{12} and accordingly obtained the partial rate of muon capture to all the bound states of B_5^{12}. He further gave a qualitative argument in favor of the view that most of the muon capture transitions to the bound states of B_5^{12} actually go to the ground state of B_5^{12}. As a result he identified his *observed* partial muon capture rate with the rate from C_6^{12} to the B_5^{12} ground state. Godfrey then established an approximate theoretical relation between the nuclear matrix elements for muon capture and for beta decay between the ground states of C_6^{12} and B_5^{12}. By using this relation and comparing his observed muon capture rate with the known B_5^{12} decay rate he concluded that the Gamow-Teller coupling constants in muon capture and in beta decay are approximately equal.

Fujii and Primakoff[1] recently re-examined and refined the relation between the nuclear matrix elements for muon capture and for beta decay between the ground states of C_6^{12} and B_5^{12}, and also extended the argument to the calculation of the ground state to ground-state partial muon capture rates in the reactions,

$$\mu^-+\mathrm{Li}_3^6 \rightarrow \nu+\mathrm{He}_2^6,$$
$$\mu^-+\mathrm{He}_2^3 \rightarrow \nu+\mathrm{H}_1^3.$$

Using the expressions for the muon capture transition matrix elements, $\mathrm{M.E.}^{(\mu)}$, given in Eqs. (2a) and (2b) and the analogous expressions for the corresponding beta-decay transition matrix elements, $\mathrm{M.E.}^{(\beta)}$,

$$|\mathrm{M.E.}^{(\beta)}|^2 = \frac{1}{(2\pi)^6} F_{ba}(Z,E_e)\,|\mathrm{M.E._{nucl}}^{(\beta)}(b\rightarrow a)|^2, \quad (21a)$$

$$|\mathrm{M.E._{nucl}}^{(\beta)}(b\rightarrow a)|^2 = (g_V{}^{(\beta)})^2|\langle a|\textstyle\sum_i \tau_i{}^{(+)}|b\rangle|^2 + (g_A{}^{(\beta)})^2|\langle a|\textstyle\sum_i \tau_i{}^{(+)}\boldsymbol{\sigma}_i|b\rangle|^2. \quad (21b)$$

Fujii and Primakoff obtained an expression for the ratio of transition rates of muon capture from $|a\rangle$ to $|b\rangle$ and beta decay from $|b\rangle$ to $|a\rangle$, $\Lambda^{(\mu)}(a\rightarrow b)/\Lambda^{\beta}(b\rightarrow a)$,

$$\frac{\Lambda^{(\mu)}(a\rightarrow b)}{\Lambda^{(\beta)}(b\rightarrow a)} = \left[\pi(\eta_{ba})^2\frac{Z^3 m_\mu{}^5}{(137)^3}\frac{1}{f_{ba}}\right]\times\left(\frac{2J_b+1}{2J_a+1}\right) \times\left[\frac{\int\frac{d\nu_1}{4\pi}\sum_{M_b,M_a}|\mathrm{M.E._{nucl}}^{(\mu)}(a\rightarrow b)|^2}{\sum_{M_b,M_a}|\mathrm{M.E._{nucl}}^{(\beta)}(b\rightarrow a)|^2}\right] \equiv k\times\left(\frac{2J_b+1}{2J_a+1}\right)\times X, \quad (22a)$$

where

$$f_{ba}=\int_1^{(E_e)_{\max}} = F_{ba}(Z,E_e)((E_e)_{\max}-E_e)^2 \times E_e((E_e)^2-1)^{\frac{1}{2}}dE_e \quad (22b)$$

$F_{ba}(Z,E)$ = Fermi function.

Fujii and Primakoff then calculated, employing appropriate approximations for $|a\rangle$, $|b\rangle$, the ratio of the nuclear matrix elements for muon capture and for beta decay—X in Eq. (22a)—as:

$$X(\mathrm{He}_2^3 \leftrightarrows \mathrm{H}_1^3)=0.791$$
$$X(\mathrm{Li}_3^6 \leftrightarrows \mathrm{He}_2^6)=0.619 \quad (23)$$
$$X(\mathrm{C}_6^{12} \leftrightarrows \mathrm{B}_5^{12})=0.612$$

and, using also known values for η_{ba} [Eq. (3c)] and $\Lambda^{(\beta)}(b\rightarrow a)/f_{ba}=\ln 2/[f_{ba}\tau_{\frac{1}{2}}(b\rightarrow a)]$, found, from Eqs. (23) and (22a),

$$\Lambda^{(\mu)}(\mathrm{He}_2^3\rightarrow \mathrm{H}_1^3)=1.46\times10^3\ \mathrm{sec}^{-1},$$
$$\Lambda^{(\mu)}(\mathrm{Li}_3^6\rightarrow \mathrm{He}_2^6)=1.79\times10^3\ \mathrm{sec}^{-1}, \quad (24)$$
$$\Lambda^{(\mu)}(\mathrm{C}_6^{12}\rightarrow \mathrm{B}_5^{12})=7.86\times10^3\ \mathrm{sec}^{-1},$$

with an over-all uncertainty of some 10 to 15%. A calculation by Wolfenstein,[15] based on general assumptions entirely similar to those of Fujii and Primakoff, yields $\Lambda^{(\mu)}(\mathrm{C}_6^{12}\rightarrow \mathrm{B}_5^{12})=7.4\times10^3\ \mathrm{sec}^{-1}$.

As regards experimental values of $\Lambda^{(\mu)}(a\rightarrow b)$, data are at present available only in the $\mathrm{C}_6^{12}\rightarrow \mathrm{B}_5^{12}$ case and are:

$$[\Lambda^{(\mu)}(\mathrm{C}_6^{12}\rightarrow \mathrm{B}_5^{12})]_{\mathrm{exper}} = \begin{cases} (9.05\pm0.95)\times10^3\ \mathrm{sec}^{-1},{}^{16} \\ (9.18\pm0.5)\times10^3\ \mathrm{sec}^{-1},{}^{17} \\ (6.6\pm1.1)\times10^3\ \mathrm{sec}^{-1},{}^{18} \\ (6.8\pm1.5)\times10^3\ \mathrm{sec}^{-1},{}^{19} \\ (5.9\pm1.5)\times10^3\ \mathrm{sec}^{-1}.{}^{14} \end{cases} \quad (25)$$

---

[15] L. Wolfenstein, Conference on Weak Interactions, Gatlinburg, Tennessee (1958, to be published).

[16] Argo, Harrison, Kruse, and McGuire, Conference on Weak Interactions, Gatlinburg, Tennessee (1958), and preprint (to be published).

[17] Burgman, Fischer, Leontic, Lundby, Meunier, Stroot, and Teja, Phys. Rev. Letters 1, 469 (1958).

[18] Fetkovich, Fields, and McIlwain, Conference on Weak Interactions, Gatlinburg, Tennessee (1958).

[19] Love, Marder, Nadelhaft, Siegel, and Taylor, Conference on Weak Interactions, Gatlinburg, Tennessee (1958, to be published), R. Siegel (private communication).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

TABLE II.

$\Lambda^{(\mu)}(C_6^{12}\to B_5^{12})$ in $sec^{-1}\times10^{-3}$	Terms $\sim(\mu_p-\mu_n)$	$g_P^{(\mu)}/g_A^{(\beta)}$
7.86	included	+8
6.34	omitted	+8
11.80	included	−8
10.25	omitted	−8

It is thus clear, from Eqs. (24) and (25), that the theoretical value of $\Lambda^{(\mu)}(C_6^{12}\to B_5^{12})$ agrees, within the over-all theoretical and experimental uncertainties, with the corresponding experimental value. This agreement offers further support for the validity of our effective Hamiltonian $H_{\text{eff}}^{(\mu)}$ [Eqs. (1a) to (1c)].

It may be of interest in concluding this section to append a table—Table II—giving the theoretical values of $\Lambda^{(\mu)}(C_6^{12}\to B_5^{12})$ with the assumption of the conserved vector current abandoned, i.e., with anomalous magnetic moment contributions $[\sim(\mu_p-\mu_n)]$ to $G_A^{(\mu)}$, $G_P^{(\mu)}$ [Eq. (1b)] omitted and with $g_P^{(\mu)}$ {within $G_P^{(\mu)}$ [Eq. (1c)]} taken as $-8g_A^{(\beta)}$ as well as $+8g_A^{(\beta)}$. Such an ambiguity in the sign of $g_P^{(\mu)}/g_A^{(\beta)}$ may be contemplated since, without a sufficiently detailed theory of the $\pi^+\to\mu^++\nu$ process,[3] one can fix, on the basis of the known π^+ lifetime, only the square of the effective coupling constant for the $\mu^-+\pi^+\to\nu$ "step" [3,4] in the $\mu^-+p\to\mu^-+\pi^++n\to\nu+n$ "two-step" process. We find from Eqs. (2a), (2b), (21a) to (22b), (1b), and (1c), the numerical results given in Table II. These results, together with the values of $[\Lambda^{(\mu)}(C_6^{12}\to B_5^{12})]_{\text{exper}}$ in Eq. (25), indicate that (1) assumption of $g_P^{(\mu)}/g_A^{(\beta)}=-8$, which is inconsistent with the usually accepted detailed theory of the $\pi^+\to\mu^++\nu$ process which involves the dominance of the $p+\bar{n}$ "intermediate state",[3] yields values of $\Lambda^{(\mu)}(C_6^{12}\to B_5^{12})$ which fit experiment less well than values of $\Lambda^{(\mu)}(C_6^{12}\to B_5^{12})$ calculated with the assumption of $g_P^{(\mu)}/g_A^{(\beta)}=+8$; (2) if future observations uphold the first pair of experimental values in Eq. (25), use of $g_P^{(\mu)}/g_A^{(\beta)}=+8$ will, for agreement between experiment and theory of $\Lambda^{(\mu)}(C_6^{12}\to B_6^{12})$, require inclusion of the anomalous magnetic moment contributions to $G_A^{(\mu)}$, $G_P^{(\mu)}$ and so support the assumption of a conserved vector current. This last conclusion agrees with that reached at the end of Sec. 3 on the basis of a comparison of experiment and theory for the total muon capture rate, $\Lambda^{(\mu)}(a)$.

5. "HYPERFINE" EFFECT IN MUON CAPTURE RATE—MUON CAPTURE IN HYDROGEN

The total muon capture rate, $\Lambda^{(\mu)}(a)$, calculated in Sec. 2 or the partial muon capture rate, $\Lambda^{(\mu)}(a\to b)$, calculated in Sec. 4 are actually appropriate averages of the total or partial muon capture rates from the two individual hyperfine states of the parent mu-mesic atom, $\Lambda^{(\mu)}(J_a\pm\frac{1}{2};a)$ or $\Lambda^{(\mu)}(J_a\pm\frac{1}{2};a\to b)$, e.g.,

$$\Lambda^{(\mu)}(a)=\frac{2J_a+2}{4J_a+2}\Lambda^{(\mu)}(J_a+\tfrac{1}{2};a)+\frac{2J_a}{4J_a+2}\Lambda^{(\mu)}(J_a-\tfrac{1}{2};a). \tag{26}$$

Equation (26) represents such an appropriate average for $\Lambda^{(\mu)}(a)$ in terms of $\Lambda^{(\mu)}(J_a\pm\frac{1}{2};a)$; this average is "incoherent" and is weighted only according to the degeneracies of the two hyperfine states involved since (1) the energy difference, $\delta\epsilon_a$, between these two hyperfine states is much greater than their width, i.e.,

$$\delta\epsilon_a\gg\hbar[\Lambda_{\text{decay}}{}^{(\mu)}+\Lambda^{(\mu)}(J_a+\tfrac{1}{2};a)];$$

$$\delta\epsilon_a\gg\hbar[\Lambda_{\text{decay}}{}^{(\mu)}+\Lambda^{(\mu)}(J_a-\tfrac{1}{2};a)];$$

and (2) the rate of conversion from one of these states to the other is (with the exception of hydrogen and deuterium) sufficiently smaller than $\Lambda_{\text{decay}}{}^{(\mu)}+\Lambda^{(\mu)}(J_a\pm\frac{1}{2};a)$ for the various mu-mesic atoms of interest.[20] The physical reason for the difference between $\Lambda^{(\mu)}(J_a+\frac{1}{2};a)$ and $\Lambda^{(\mu)}(J_a-\frac{1}{2};a)$ arises, as discussed by Bernstein, Lee, Yang, and Primakoff [21] (B.L.Y. and P.), from the combined action of the following three effects (1) the correlation between the spin of the muon, $\frac{1}{2}\boldsymbol{\sigma}$, and the spin of the parent nucleus $\mathbf{J}_a$ is different in the two hyperfine states $F_a^{(\pm)}=J_a\pm\frac{1}{2}$; (2) there is, in general, a correlation between the spin $\frac{1}{2}\boldsymbol{\sigma}_i$ of the proton that captures the muon and $\mathbf{J}_a$; (3) the capture rate of the muon by the proton depends on their relative spin orientation via the terms $G_A^{(\mu)}\boldsymbol{\sigma}\cdot\boldsymbol{\sigma}_i$ and $G_P^{(\mu)}\boldsymbol{\sigma}\cdot\boldsymbol{\nu}_1\boldsymbol{\sigma}_i\cdot\boldsymbol{\nu}_1$ in the effective Hamiltonian $H_{\text{eff}}^{(\mu)}$ of Eqs. (1a) to (1c).

The total muon capture rates from the two individual hyperfine states $F_a^{(\pm)}=J_a\pm\frac{1}{2}$, $\Lambda^{(\mu)}(J_a\pm\frac{1}{2};a)$, can be calculated on the basis of the $H_{\text{eff}}^{(\mu)}$ of Eqs. (1a) to (1c) by the closure approximation method described in Sec. 2. By using a procedure similar to that involved in the derivation of Eqs. (3a) to (8) we find that

---

[20] Conversion from the energetically higher to the energetically lower of the two hyperfine mu-mesic atom states occurs (1) via collisions with atoms of the muon moderating medium—this is important only for the electrically neutral and hence mobile hydrogen and deuterium mu-mesic atoms—see below; (2) via spontaneous magnetic dipole radiation; (3) via Auger electron ejection. The rates of (2) and (3) may be readily estimated and are, respectively:

$$R_{(2)}\approx10\left(\frac{1}{137}\right)^{13}(Z_{\text{eff}})^9\left(\frac{m_\mu}{m_p}\right)^3\frac{m_\mu c^2}{\hbar}$$

and

$$R_{(3)}\approx10\left(\frac{1}{137}\right)^{6}Z\left(\frac{m_e}{m_\mu}\right)^3\frac{m_\mu c^2}{\hbar},$$

so that for example, $R_{(2)}\approx1\times10^3$ sec^{-1}, $R_{(3)}\approx4\times10^5$ sec^{-1} for $Al_{13}{}^{27}$; on the other hand, the value of

$$\{\Lambda_{\text{decay}}{}^{(\mu)}+\Lambda^{(\mu)}(J_a\pm\tfrac{1}{2};a)\}$$

for $Al_{13}{}^{27}$ is $\approx1.2\times10^6$ sec^{-1}.

[21] Bernstein, Lee, Yang, and Primakoff, Phys. Rev. 111, 313 (1958).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

$$\frac{\Lambda^{(\mu)}(J_a\pm\tfrac{1}{2};a)}{\Lambda^{(\mu)}(a)}-1\cong\pm(J_a+\tfrac{1}{2}\pm\tfrac{1}{2})^{-1}\frac{b^{(\mu)}}{a^{(\mu)}}\frac{\langle a|2\mathbf{J}\cdot\mathbf{S}_{\rm pr}|a\rangle}{Z}$$

$$\times\left\{\frac{\xi_a-\left[\frac{1}{2}+\frac{1}{2}\frac{d^{(\mu)}}{b^{(\mu)}}\frac{\langle a|2\mathbf{J}\cdot\mathbf{S}_{\rm neu}|a\rangle}{\langle a|2\mathbf{J}\cdot\mathbf{S}_{\rm pr}|a\rangle}-\frac{1}{2}\frac{\langle a|2\mathbf{J}\cdot\mathbf{K}|a\rangle}{\langle a|2\mathbf{J}\cdot\mathbf{S}_{\rm pr}|a\rangle}\right]\alpha_a{}'^{(+)}-\left[\frac{A-Z}{2}+\frac{Z}{2}\frac{d^{(\mu)}}{b^{(\mu)}}\frac{\langle a|2\mathbf{J}\cdot\mathbf{S}_{\rm neu}|a\rangle}{\langle a|2\mathbf{J}\cdot\mathbf{S}_{\rm pr}|a\rangle}\right]\alpha_a{}'^{(-)}}{1-\left[\frac{A}{2Z}-\frac{\gamma_a{}^{(+)}}{Za^{(\mu)}}\right]\alpha_a{}^{(+)}-\left[\frac{A-Z}{2}-\frac{\gamma_a{}^{(-)}}{Za^{(\mu)}}\right]\alpha_a{}^{(-)}}\right\},\tag{27a}$$

where

$$b^{(\mu)}\equiv 2(G_V{}^{(\mu)}G_A{}^{(\mu)}-\tfrac{1}{3}G_V{}^{(\mu)}G_P{}^{(\mu)})-2[(G_A{}^{(\mu)})^2-\tfrac{2}{3}G_A{}^{(\mu)}G_P{}^{(\mu)}],$$

$$d^{(\mu)}\equiv 2(G_V{}^{(\mu)}G_A{}^{(\mu)}-\tfrac{1}{3}G_V{}^{(\mu)}G_P{}^{(\mu)})+2[(G_A{}^{(\mu)})^2-\tfrac{2}{3}G_A{}^{(\mu)}G_P{}^{(\mu)}],$$

$$a^{(\mu)}\equiv(G_V{}^{(\mu)})^2+3(\Gamma_A{}^{(\mu)})^2=(G_V{}^{(\mu)})^2+3(G_A{}^{\mu})^2+(G_P{}^{(\mu)})^2-2G_A{}^{(\mu)}G_P{}^{(\mu)},$$

$$\mathbf{K}\equiv(\mathbf{Y}^{(1)}T^{(1)}+\mathbf{Y}^{(2)}T^{(2)})-\left(\frac{b^{(\mu)}-d^{(\mu)}}{2b^{(\mu)}}\right)(\mathbf{Y}^{(1)}T^{(1)}+\mathbf{Y}^{(2)}T^{(2)}+\mathbf{Y}^{(1)}\times\mathbf{Y}^{(2)}),$$

$$\xi_a\equiv\left(\int|\varphi(r)|^2\mathfrak{D}_a{}'(r)d\mathbf{r}\right)\Big/\left(\int|\varphi(r)|^2\mathfrak{D}_a(r)d\mathbf{r}\right),$$

$$\alpha_a{}'^{(\pm)}\equiv\left[\int\int\frac{\sin(\langle\nu\rangle_a|\mathbf{r}-\mathbf{r}'|)}{(\langle\nu\rangle_a|\mathbf{r}-\mathbf{r}'|)}\varphi^*(r)\varphi(r')\tfrac{1}{2}(F_a{}'^{(+)}(\mathbf{r},\mathbf{r}')\pm F_a{}'^{(-)}(\mathbf{r},\mathbf{r}'))d\mathbf{r}d\mathbf{r}'\right]\Big/\left[\int|\varphi(r)|^2\mathfrak{D}_a{}'(r)d\mathbf{r}\right],\tag{27b}$$

$$\mathfrak{D}_a{}'(r)\equiv\frac{1}{2J_a+1}\sum_{M_a}\frac{\langle E_a,J_a,M_a|\mathbf{J}\cdot\sum_i[(1+\tau_i{}^{(3)})/2]\boldsymbol{\sigma}_i\delta(\mathbf{r}-\mathbf{r}_i)|E_a,J_a,M_a\rangle}{\langle E_a,J_a,M_a|\mathbf{J}\cdot\sum_i[(1+\tau_i{}^{(3)})/2]\boldsymbol{\sigma}_i|E_a,J_a,M_a\rangle},$$

$$F_a{}'^{(\pm)}(\mathbf{r},\mathbf{r}')\equiv\frac{\langle a|\mathbf{J}\cdot\sum'_{i,j}[\frac{1}{4}(\boldsymbol{\tau}_i\cdot\boldsymbol{\tau}_j-\tau_i{}^{(3)}\tau_j{}^{(3)})(b^{(\mu)}+d^{(\mu)})(\boldsymbol{\sigma}_i+\boldsymbol{\sigma}_j)-\frac{1}{4}(\boldsymbol{\tau}_i\times\boldsymbol{\tau}_j)^{(3)}(b^{(\mu)}-d^{(\mu)})(\boldsymbol{\sigma}_i\times\boldsymbol{\sigma}_j)]\times[\delta(\mathbf{r}-\mathbf{r}_i)\delta(\mathbf{r}'-\mathbf{r}_j)](1+P_{ij})/2|a\rangle}{\langle a|\mathbf{J}\cdot\sum'_{i,j}[\frac{1}{4}(\boldsymbol{\tau}_i\cdot\boldsymbol{\tau}_j-\tau_i{}^{(3)}\tau_j{}^{(3)})(b^{(\mu)}+d^{(\mu)})(\boldsymbol{\sigma}_i+\boldsymbol{\sigma}_j)-\frac{1}{4}(\boldsymbol{\tau}_i\times\boldsymbol{\tau}_j)^{(3)}(b^{(\mu)}-d^{(\mu)}(\boldsymbol{\sigma}_i\times\boldsymbol{\sigma}_j)]\times(1+P_{ij})/2|a\rangle},$$

$$\alpha_a{}^{(\pm)}\equiv\left[\int\int\frac{\sin(\langle\nu\rangle_a|\mathbf{r}-\mathbf{r}'|)}{(\langle\nu\rangle_a|\mathbf{r}-\mathbf{r}'|)}\varphi^*(r)\varphi(r')\tfrac{1}{2}(F_a{}^{(+)}(\mathbf{r},\mathbf{r}')\pm F_a{}^{(-)}(\mathbf{r},\mathbf{r}'))d\mathbf{r}d\mathbf{r}'\right]\Big/\left[\int|\varphi(r)|^2\mathfrak{D}_a(r)d\mathbf{r}\right],$$

$$\gamma_a{}^{(+)}\equiv\langle a|(G_V{}^{(\mu)})^2((\mathbf{T})^2-(T^{(3)})^2)+(\Gamma_A{}^{(\mu)})^2((\mathbf{Y}^{(1)})^2+(\mathbf{Y}^{(2)})^2)|a\rangle,$$

$$\gamma_a{}^{(-)}\equiv((G_V{}^{(\mu)})^2-(\Gamma_A{}^{(\mu)})^2)\langle a|(\mathbf{S}_{\rm pr})^2+(\mathbf{S}_{\rm neu})^2-(\mathbf{S}_{\rm pr}+\mathbf{S}_{\rm neu})^2|a\rangle.$$

We now apply Eq. (27a) to the case of an odd Z, odd A nucleus so that $\langle a|2\mathbf{J}\cdot\mathbf{S}_{\rm neu}|a\rangle=0$. Considering also the heavier nuclei, $Z>6$, $A>12$, and hence dropping terms $\sim 1/Z$ within the curly bracket of Eq. (27a), we have:

$$\frac{\Lambda^{(\mu)}(J_a\pm\tfrac{1}{2};a)}{\Lambda^{(\mu)}(a)}-1\cong\pm(J_a+\tfrac{1}{2}\pm\tfrac{1}{2})^{-1}\frac{b^{(\mu)}}{a^{(\mu)}}\frac{\langle a|2\mathbf{J}\cdot\mathbf{S}_{\rm pr}|a\rangle}{Z}\left\{\frac{\xi_a-\left[\frac{1}{2}-\frac{1}{2}\frac{\langle a|2\mathbf{J}\cdot\mathbf{K}|a\rangle}{\langle a|2\mathbf{J}\cdot\mathbf{S}_{\rm pr}|a\rangle}\right]\alpha_a{}'^{(+)}-\left[\frac{A-Z}{2}\right]\alpha_a{}'^{(-)}}{1-[A/2Z]\alpha_a{}^{(+)}-[(A-Z)/2]\alpha_a{}^{(-)}}\right\}\tag{28a}$$

so that

$$\frac{\Lambda^{(\mu)}(J_a+\tfrac{1}{2};a)-\Lambda^{(\mu)}(J_a-\tfrac{1}{2};a)}{\Lambda^{(\mu)}(a)}\cong\frac{2J_a+1}{J_a(J_a+1)}\frac{b^{(\mu)}}{a^{(\mu)}}\frac{\langle a|2\mathbf{J}\cdot\mathbf{S}_{\rm pr}|a\rangle}{Z}\left\{\frac{\xi_a-\left[\frac{1}{2}-\frac{1}{2}\frac{\langle a|2\mathbf{J}\cdot\mathbf{K}|a\rangle}{\langle a|2\mathbf{J}\cdot\mathbf{S}_{\rm pr}|a\rangle}\right]\alpha_a{}'^{(+)}-\left[\frac{A-Z}{2}\right]\alpha_a{}'^{(-)}}{1-[A/2Z]\alpha_a{}^{(+)}-[(A-Z)/2]\alpha_a{}^{(-)}}\right\}.\tag{28b}$$

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

The quantity in the curly brackets in Eq. (28a) or (28b) is the ratio of the exclusion principle inhibition factors for $\Lambda^{(\mu)}(J_a+\frac{1}{2};a)-\Lambda^{(\mu)}(J_a-\frac{1}{2};a)$ and for $\Lambda^{(\mu)}(a)$ and should be close to unity. Assuming in addition that for the purpose of calculating $\langle a|2\mathbf{J}\cdot\mathbf{S}_{\rm pr}|a\rangle$ one can visualize the odd Z, odd A nucleus as consisting of an "outside" proton with orbital angular momentum $\mathbf{L}_a$ moving about a spinless "core," one has

$$\langle a|2\mathbf{J}\cdot\mathbf{S}_{\rm pr}|a\rangle=J_a(J_a+1)-L_a(L_a+1)+\tfrac{3}{4}=\begin{matrix}J_a+1:J_a=L_a+\frac{1}{2}\\-J_a\ :J_a=L_a-\frac{1}{2}\end{matrix} \tag{29}$$

so that, substituting into Eq. (28b),

$$\frac{\Lambda^{(\mu)}(J_a+\frac{1}{2};a)-\Lambda^{(\mu)}(J_a-\frac{1}{2};a)}{\Lambda^{(\mu)}(a)}\cong\frac{b^{(\mu)}}{a^{(\mu)}}\frac{1}{Z}\times\begin{bmatrix}(2J_a+1)/J_a\\-(2J_a+1)/(J_a+1)\end{bmatrix}\begin{matrix}:J_a=L_a+\frac{1}{2}\\:J_a=L_a-\frac{1}{2}\end{matrix} \tag{30}$$

in essential agreement with the result for

$$(\Lambda^{(\mu)}(J_a+\tfrac{1}{2};a)-\Lambda^{(\mu)}(J_a-\tfrac{1}{2};a))/\Lambda^{(\mu)}(a)$$

obtained in another way by B.L.Y. and P.[21] The combination of coupling constants in Eq. (30), $b^{(\mu)}/a^{(\mu)}$, has the numerical value, using Eqs. (27b), (1b), (1c), and (3d), and $(g_A{}^{(\beta)}/g_V{}^{(\beta)})=-1.21$,[11]

$$\frac{b^{(\mu)}}{a^{(\mu)}}=-0.945. \tag{31}$$

As an example of the magnitude of the hyperfine effect we calculate the quantity

$$(\Lambda^{(\mu)}(J_a+\tfrac{1}{2};a)-\Lambda^{(\mu)}(J_a-\tfrac{1}{2};a))/\Lambda^{(\mu)}(a)$$ from Eqs. (30) and (31) for $Al_{13}{}^{27}$. Here, $Z=13$, $J_a=\frac{5}{2}$, $L_a=2$, so that we have

$$\frac{\Lambda^{(\mu)}(J_a+\frac{1}{2};a)-\Lambda^{(\mu)}(J_a-\frac{1}{2};a)}{\Lambda^{(\mu)}(a)}\cong-0.17, \tag{32}$$

which appears sufficiently large to be observed with available experimental techniques.

We now apply Eqs. (27a) and (27b) to the case of muon capture in hydrogen. Here

$$\langle a|2\mathbf{J}\cdot\mathbf{S}_{\rm neu}|a\rangle=0;\quad\langle a|2\mathbf{J}\cdot\mathbf{S}_{\rm pr}|a\rangle=2\cdot\tfrac{3}{4};$$
$$\langle a|2\mathbf{J}\cdot\mathbf{K}|a\rangle=2\cdot\tfrac{3}{4};\quad\langle a|(\mathbf{T})^2-(T^{(3)})^2|a\rangle=\tfrac{1}{2};$$
$$\langle a|(\mathbf{Y}^{(1)})^2+(\mathbf{Y}^{(2)})^2|a\rangle=\tfrac{3}{2};$$
$$\langle a|(\mathbf{S}_{\rm pr})^2+(\mathbf{S}_{\rm neu})^2-(\mathbf{S}_{\rm pr}+\mathbf{S}_{\rm neu})^2|a\rangle=0;$$
$$b^{(\mu)}/a^{(\mu)}=-0.922$$

[from Eqs. (27b), (1b), (1c), (2c), and $(g_A{}^{(\beta)}/g_V{}^{(\beta)})=-1.21$ [11]], $\Lambda^{(\mu)}(H_1{}^1)=169$ sec^{-1} [Eq. (17)], so that

$$\Lambda^{(\mu)}(\tfrac{1}{2}+\tfrac{1}{2};H_1{}^1)=\Lambda^{(\mu)}(H_1{}^1)\left(1+\frac{b^{(\mu)}}{a^{(\mu)}}\right)=13\text{ sec}^{-1}, \tag{33a}$$

$$\Lambda^{(\mu)}(\tfrac{1}{2}-\tfrac{1}{2};H_1{}^1)=\Lambda^{(\mu)}(H_1{}^1)\left(1-\frac{3b^{(\mu)}}{a^{(\mu)}}\right)=636\text{ sec}^{-1}. \tag{33b}$$

Thus the hyperfine effect is enormous—

$$(\Lambda^{(\mu)}(\tfrac{1}{2}-\tfrac{1}{2};H_1{}^1)/\Lambda^{(\mu)}(\tfrac{1}{2}+\tfrac{1}{2};H_1{}^1))\cong50.$$

Muon capture in hydrogen is also unique in that within a time $\approx[p\mu^-]$ mu-mesic atom mean life $\tau^{\pm}$

$$(\tau^{\pm}=\{\Lambda_{\rm decay}{}^{(\mu)}+\Lambda^{(\mu)}(\tfrac{1}{2}\pm\tfrac{1}{2};H_1{}^1)\}^{-1}\approx1/\Lambda_{\rm decay}{}^{(\mu)}=2.21\times10^{-6}\text{ sec})$$

there is, under certain circumstances, a high probability of conversion from the energetically higher triplet state with $F_a{}^{(+)}=\frac{1}{2}+\frac{1}{2}$ to the energetically lower singlet state with $F_a{}^{(-)}=\frac{1}{2}-\frac{1}{2}$. This is a consequence of the fact that, as pointed out by Gershtein and Zeldovich,[22] the $[p\mu^-]$ mu-mesic is electrically neutral and so wanders through the hydrogen gas or liquid making collisions with the $(H_1{}^1)_2$ molecules—these collisions occasionally result in "exchange" between the proton in the incident $[p\mu^-]$ mu-mesic atom with one direction of spin and a proton in the target $(H_1{}^1)_2$ molecule with the opposite direction of spin, with the net result that the $[p\mu^-]$ is converted from the triplet state to the singlet state. The rate of such a collisional conversion process is estimated by Gershtein[22] to be $\approx5\times10^9$ sec^{-1} at an $(H_1{}^1)_2$ molecule number density, $\mathfrak{N}/V$, of 2×10^{22}/cm^3—thus at $\mathfrak{N}/V$ less than, say, 2×10^{17}/cm^3 there is no appreciable collisional triplet to singlet conversion and the muon capture rate is, from Eqs. (33a), (33b), and (26),

$$\tfrac{3}{4}\Lambda^{(\mu)}(\tfrac{1}{2}+\tfrac{1}{2};H_1{}^1)+\tfrac{1}{4}\Lambda^{(\mu)}(\tfrac{1}{2}-\tfrac{1}{2};H_1{}^1)=\Lambda^{(\mu)}(H_1{}^1)=169\text{ sec}^{-1}.$$

With increasing $\mathfrak{N}/V$ the collisional triplet to singlet conversion becomes progressively more important and at $\mathfrak{N}/V\approx10^{20}$/cm^3 practically all the $[p\mu^-]$ mu-mesic atoms are in the singlet state at the instant of muon decay or capture—the corresponding muon capture rate is, from Eq. (33b), $\Lambda^{(\mu)}(\frac{1}{2}-\frac{1}{2};H_1{}^1)=636$ sec^{-1}. At still higher $\mathfrak{N}/V$, e.g., at $\mathfrak{N}/V$ of the order of those in liquid hydrogen, formation of mu-mesic hydrogen molecule ions, $[p\mu^-p]$, becomes dominant[23] and it is expected that most of the muons will be found at the instant of their decay or capture in the lowest Bohr orbit of a

[22] S. S. Gershtein, J. Exptl. Theoret. Phys. U.S.S.R.) **34**, 463 (1958); *ibid.* **34**, 993 (1958).

[23] Y. B. Zeldovich, Doklady Akad. Nauk (S.S.S.R.) **95**, 493 (1954), and J. Exptl. Theoret. Phys. (U.S.S.R.) **33**, 310 (1957); J. D. Jackson, Phys. Rev. **106**, 330 (1957); T. H. R. Skyrme, Phil. Mag. **2**, 910 (1957); Y. B. Zeldovich and S. S. Gershtein, J. Exptl. Theoret. Phys. (U.S.S.R.) **35**, 649 (1958); Cohen, Judd, and Riddell, Phys. Rev. **110**, 1471 (1958).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

$[p\mu^- p]$. Any such $[p\mu^- p]$ must have a total spin angular momentum of $\frac{1}{2}$ (since one of its parents is a singlet $[p\mu^-]$) and so possesses a muon spin $\boldsymbol{\sigma}$—proton spin $\boldsymbol{\sigma}_1$, $\boldsymbol{\sigma}_2$ configuration intermediate between that in a triplet and that in a singlet $[p\mu_-]$—in fact

$$\langle(\mathbf{F})^2\rangle_a = \tfrac{3}{4} = \langle[\tfrac{1}{2}\boldsymbol{\sigma} + (\tfrac{1}{2}\boldsymbol{\sigma}_1 + \tfrac{1}{2}\boldsymbol{\sigma}_2)]^2\rangle_a$$
$$= \tfrac{3}{4} + \langle(\tfrac{1}{2}\boldsymbol{\sigma}_1 + \tfrac{1}{2}\boldsymbol{\sigma}_2)^2\rangle_a + 2\langle(\tfrac{1}{2}\boldsymbol{\sigma})\cdot(\tfrac{1}{2}\boldsymbol{\sigma}_1 + \tfrac{1}{2}\boldsymbol{\sigma}_2)\rangle_a$$

so that

$$\langle\boldsymbol{\sigma}\cdot\boldsymbol{\sigma}_1\rangle_a = \langle\boldsymbol{\sigma}\cdot\boldsymbol{\sigma}_2\rangle_a = -\langle(\tfrac{1}{2}\boldsymbol{\sigma}_1 + \tfrac{1}{2}\boldsymbol{\sigma}_2)^2\rangle_a = 0, -2$$

in *para*, *ortho* $[p\mu^- p]$ while $\langle\boldsymbol{\sigma}\cdot\boldsymbol{\sigma}_1\rangle_a = 1, -3$ in triplet, singlet $[p\mu^-]$. Further, since from Eqs. (33a) and (33b),

$$\Lambda^{(\mu)}(\tfrac{1}{2}\pm\tfrac{1}{2}; \mathrm{H}_1{}^1) \sim \left(1 + \langle\boldsymbol{\sigma}\cdot\boldsymbol{\sigma}_1\rangle_{a;\pm}\frac{b^{(\mu)}}{a^{(\mu)}}\right)$$
$$= \left(1 + \begin{pmatrix}1\\-3\end{pmatrix}\times\frac{b^{(\mu)}}{a^{(\mu)}}\right)$$

the muon capture rate in *para* $[p\mu^- p]$, *ortho* $[p\mu^- p]$ will, in an analogous fashion, be

$$\sim\left(1 + \langle\boldsymbol{\sigma}\cdot\boldsymbol{\sigma}_1\rangle_{a;\,para,\,ortho}\frac{b^{(\mu)}}{a^{(\mu)}}\right) = \left(1 + \begin{pmatrix}0\\-2\end{pmatrix}\times\frac{b^{(\mu)}}{a^{(\mu)}}\right).$$

It follows that at high $\mathfrak{N}/V$ the muon capture rate has the form

$$x_1\left(\frac{\mathfrak{N}}{V}\right)\left[\Lambda^{(\mu)}(\mathrm{H}_1{}^1)\left(1 - \frac{3b^{(\mu)}}{a^{(\mu)}}\right)\right]$$
$$+ x_2\left(\frac{\mathfrak{N}}{V}\right)[(2\gamma)\Lambda^{(\mu)}(\mathrm{H}_1{}^1)]$$
$$+ x_3\left(\frac{\mathfrak{N}}{V}\right)\left[(2\gamma)\Lambda^{(\mu)}(\mathrm{H}_1{}^1)\left(1 - \frac{2b^{(\mu)}}{a^{(\mu)}}\right)\right]$$
$$= \left\{x_1\left(\frac{\mathfrak{N}}{V}\right)[636] + x_2\left(\frac{\mathfrak{N}}{V}\right)[(2\gamma)(169)]\right.$$
$$\left. + x_3\left(\frac{\mathfrak{N}}{V}\right)[(2\gamma)(480)]\right\}\ \mathrm{sec}^{-1}, \quad (34)$$

where $x_1(\mathfrak{N}/V)$, $x_2(\mathfrak{N}/V)$, $x_3(\mathfrak{N}/V)$, are the fractions of muons found in singlet $[p\mu^-]$, *para* $[p\mu^- p]$ and *ortho* $[p\mu^- p]$ at the $\mathfrak{N}/V$ in question—$x_1(\mathfrak{N}/V) + x_2(N/V) + x_3(\mathfrak{N}/V) = 1$—and γ is the ratio of the absolute square of the muon orbital wave function at the proton position in $[p\mu^- p]$ and in $[p\mu^-]$. Equations (33a), (33b), and (34) show that the muon capture rate in hydrogen, considered as a function of $\mathfrak{N}/V$, exhibits a maximum, $\cong 636$ sec^{-1}, at an intermediate value of $\mathfrak{N}/V$; if in addition $2\gamma \cong 1$ and $(x_2(\mathfrak{N}/V)/x_3(\mathfrak{N}/V)) \gg 1$ this rate falls to values $\cong 169$ sec^{-1} at high $\mathfrak{N}/V$ as well as at very low $\mathfrak{N}/V$.

It is necessary for the physical relevance of the whole above discussion that the isotopically natural hydrogen in which the muons stop be purified of deuterium to an extent—factor ≈ 25–50 in relative abundance—sufficient to prevent any frequent occurrence of the "exchange" reaction $[p\mu^-] + d \rightarrow [d\mu^-] + p$. Any appreciable rate for such an exchange reaction with the subsequent formation of $[d\mu^- p]$ and the resultant "muon catalysis" of $d + p \rightarrow \mathrm{He}_2{}^3 + \gamma$ complicates the situation very considerably. On the other hand, a similar analysis may be given for muons stopping in pure deuterium with the major difference that at high $\mathfrak{N}/V$ the sequence of reactions:

$$[d\mu^-] + d \rightarrow [d\mu^- d]; \quad [d\mu^- d] \nearrow \mathrm{He}_2{}^3 + n + \mu^-$$
$$\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \searrow \mathrm{H}_1{}^3 + p + \mu^-$$

is anticipated with most of the muons eventually being trapped in Bohr orbits about the $\mathrm{He}_2{}^3$ nuclei.[24] Thus most muons stopping in pure liquid deuterium which are eventually captured by nuclei, are captured by nuclei of $\mathrm{He}_2{}^3$, a circumstance which permits study of muon capture in $\mathrm{He}_2{}^3$ without any anterior possession of a $\mathrm{He}_2{}^3$ target.

In concluding this section we remark again that the study of muon capture in hydrogen will not only ultimately yield the most reliable values of the effective muon—dressed nucleon coupling constants, $G_V{}^{(\mu)}$, $G_A{}^{(\mu)}$, $G_P{}^{(\mu)}$, i.e., values free of "nuclear physics" uncertainties, but will also shed light on several other interesting effects such as the collisional conversion in $[p\mu^-]$ and the formation of *para* $[p\mu^- p]$ and *ortho* $[p\mu^- p]$.

6. RADIATIVE MUON CAPTURE—TOTAL RATE AND PHOTON-NEUTRINO ANGULAR CORRELATION

We now discuss the process of radiative muon capture. Here calculations of the total rate and of the shape of the corresponding internal bremsstrahlung (I.B.) momentum spectrum have been made by Cantwell[25] for capture by light and medium-heavy nuclei ($Z/137 \ll 1$). Cantwell uses the effective Hamiltonian of Eqs. (1a) and (1b) but with all nucleon recoil corrections, i.e. all terms $\sim \nu/m_p$ omitted so that $G_V{}^{(\mu)} = g_V{}^{(\mu)}$; $G_A{}^{(\mu)} = g_A{}^{(\mu)}$; $G_P{}^{(\mu)} = 0$. The radiative capture is visualized as predominantly due to the "two-step process"

$$\mu^- + p \rightarrow \gamma + (\mu^-)' + p \rightarrow \gamma + \nu + n \quad (35)$$

and a second-order perturbation calculation, using an appropriate (free particle) Green's function to describe the virtual intermediate μ^- states, is carried out. In this way Cantwell obtains a relation for the relative

[24] J. D. Jackson, reference 23.
[25] R. M. Cantwell, Ph.D. thesis, Washington University (1956).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

probability of radiative to nonradiative muon capture

$$\frac{\sum_b \Lambda_{\rm rad}{}^{(\mu)}(a\to b;\gamma_{ba})d(\gamma_{ba}/\gamma_{ba}{}^{\rm max})d\gamma_1}{\sum_b \Lambda^{(\mu)}(a\to b)}$$

$$=\sum_b\left[\left(\frac{1}{4\pi^2}\frac{1}{137}\right)\left(\frac{\gamma_{ba}{}^{\rm max}}{m_\mu}\right)^2\left(1-\frac{\gamma_{ba}}{\gamma_{ba}{}^{\rm max}}\right)^2\left(\frac{\gamma_{ba}}{\gamma_{ba}{}^{\rm max}}\right)d\left(\frac{\gamma_{ba}}{\gamma_{ba}{}^{\rm max}}\right)d\gamma_1\right]$$

$$\times\left[\int\frac{d\nu_1}{4\pi}\{(g_V{}^{(\mu)})^2(1+\nu_1\cdot\gamma_1)|\langle b|\exp'|a\rangle|^2+(g_A{}^{(\mu)})^2(1-\nu_1\cdot\gamma_1)|\langle b|(\exp')\sigma|a\rangle|^2\right.$$

$$\left.+(g_A{}^{(\mu)})^2 2\,\mathrm{Re}(\nu_1\cdot\langle b|(\exp')\sigma|a\rangle^*\gamma_1\cdot\langle b|(\exp')\sigma|a\rangle)\right]$$

$$=\frac{}{\sum_b\left[\int\frac{d\nu_1}{4\pi}\{(g_V{}^{(\mu)})^2|\langle b|\exp|a\rangle|^2+(g_A{}^{(\mu)})^2|\langle b|(\exp)\sigma|a\rangle|^2\}\right]}, \tag{36a}$$

where

$$\exp'\equiv\sum_i\tau_i{}^{(-)}\exp[-i(\nu_{ba}{}'+\gamma_{ba})\cdot\mathbf{r}_i]\varphi(r_i);\quad(\exp')\sigma\equiv\sum_i\tau_i{}^{(-)}\exp[-i(\nu_{ba}{}'+\gamma_{ba})\cdot\mathbf{r}_i]\varphi(r_i)\sigma_i,$$
$$\exp\equiv\sum_i\tau_i{}^{(-)}\exp(-i\nu_{ba}\cdot\mathbf{r}_i)\varphi(r_i);\quad(\exp)\sigma\equiv\sum_i\tau_i{}^{(-)}\exp(-i\nu_{ba}\cdot\mathbf{r}_i)\varphi(r_i)\sigma_i. \tag{36b}$$

In Eq. (36a) the sum over b runs over all the energetically accessible states of the daughter nucleus; $\nu_{ba}{}'=\nu_{ba}{}'\nu_1$, $\gamma_{ba}=\gamma_{ba}\gamma_1$ are the momenta of the neutrino and I.B. photon emitted in the radiative capture; the energy, momentum conservation laws yield, analogously to Eqs. (2c) and (3d),

$$\nu_{ba}{}'+\gamma_{ba}\cong m_\mu\left[1-\frac{\epsilon_a}{m_\mu}-\frac{(E_b-E_a)}{m_\mu}\right]\cong\nu_{ba}. \tag{36c}$$

The first square bracket in the numerator represents an I.B. momentum spectrum of the same shape as that found by Morrison and Schiff in allowed K-electron orbital capture[26]; the second square bracket describes the dependence of the I.B. momentum spectrum on the wave functions $|a\rangle$, $|b\rangle$ of the parent and daughter nuclei—this second square bracket is however such for $\gamma_{ba}/\gamma_{ba}{}^{\rm max}=\gamma_{ba}/\nu_{ba}\cong 0$ and $\cong 1$ that the I.B. momentum spectrum has the allowed K-capture shape both near its low-energy end and its high-energy end. Thus a careful study of the shape of I.B. momentum spectrum near its high-energy end will yield $\langle\gamma\rangle_a{}^{\rm max}$, the average value of $\gamma_{ba}{}^{\rm max}$ over the states b; since $\langle\gamma\rangle_a{}^{\rm max}=\langle\nu\rangle_a$ one will in this way find empirically the average value of the neutrino energy involved in nonradiative muon capture by a particular parent nucleus and so remove the major uncertainty associated with the evaluation of the coupling constant ratio $\mathfrak{R}$ [Eq. (6b)] from the experimental data [see Sec. 3, Eqs. (19), (20), *et seq.*].

We now apply the closure approximation for the evaluation of the sums over b in Eq. (36a). By using the techniques of Eqs. (3b) to (6c), we obtain

$$\frac{\Lambda_{\rm rad}{}^{(\mu)}(a;\langle\gamma\rangle_a)dx_a d\gamma_1}{\Lambda^{(\mu)}(a)}=\left[\left(\frac{1}{4\pi^2}\frac{1}{137}\right)\left(\frac{\langle\gamma\rangle_a{}^{\rm max}}{m_\mu}\right)^2(1-x_a)^2x_a dx_a d\gamma_1\right]$$

$$\times\left\{\frac{((g_V{}^{(\mu)})^2+3(g_A{}^{(\mu)})^2)\int\frac{d\nu_1}{4\pi}(1-g_a{}'(\nu_1,\gamma_1))+((g_V{}^{(\mu)})^2-(g_A{}^{(\mu)})^2)\int\frac{d\nu_1}{4\pi}\nu_1\cdot\gamma_1(1-g_a{}''(\nu_1,\gamma_1))}{((g_V{}^{(\mu)})^2+3(g_A{}^{(\mu)})^2)(1-g_a)}\right\}, \tag{37a}$$

where g_a is the nonradiative capture exclusion principle inhibition factor given in Eq. (6c) and $g_a{}'(\nu_1,\gamma_1)$, $g_a{}''(\nu_1,\gamma_1)$ are the corresponding radiative capture inhibition factors

$$g_a{}'(\nu_1,\gamma_1)=-\frac{\langle a|\sum_{i,j}{}'[\tfrac{1}{4}(\tau_i\cdot\tau_j-\tau_i{}^{(3)}\tau_j{}^{(3)})((g_V{}^{(\mu)})^2+(g_A{}^{(\mu)})^2\sigma_i\cdot\sigma_j)][\exp\{i(\langle\nu'\rangle_a\nu_1+\langle\gamma\rangle_a\gamma_1)\cdot\mathbf{r}_{ij}\}\varphi^*(r_i)\varphi(r_j)]|a\rangle}{Z[(g_V{}^{(\mu)})^2+3(g_A{}^{(\mu)})^2]\left[\int|\varphi(\mathbf{r})|^2\mathfrak{D}_a(r)d\mathbf{r}\right]}, \tag{37b}$$

$$g_a{}''(\nu_1,\gamma_1)=-\frac{\langle a|\sum_{i,j}{}'[\tfrac{1}{4}(\tau_i\cdot\tau_j-\tau_i{}^{(3)}\tau_j{}^{(3)})((g_V{}^{(\mu)})^2-\tfrac{1}{3}(g_A{}^{(\mu)})^2\sigma_i\cdot\sigma_j)][\exp\{i(\langle\nu'\rangle_a\nu_1+\langle\gamma\rangle_a\gamma_1)\cdot\mathbf{r}_{ij}\}\varphi^*(r_i)\varphi(r_j)]|a\rangle}{Z[(g_V{}^{(\mu)})^2-(g_A{}^{(\mu)})^2]\left[\int|\varphi(\mathbf{r})|^2\mathfrak{D}_a(r)d\mathbf{r}\right]} \tag{37c}$$

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

[26] P. Morrison and L. I. Schiff, Phys. Rev. 58, 24 (1940).

with

$$\langle\nu'\rangle_a+\langle\gamma\rangle_a\cong m_\mu\left\{1-\frac{\epsilon_a}{m_\mu}-\frac{\langle\text{excitation energy of daughter nucleus } b\rangle}{m_\mu}-\frac{[(E_b)_{\text{ground state}}-E_a]}{m_\mu}\right\} \tag{37d}$$

$$x_a\equiv\frac{\langle\gamma\rangle_a}{\langle\gamma_a\rangle^{\max}}=\frac{\langle\gamma\rangle_a}{\langle\nu\rangle_a}\,.$$

As a rough first approximation to evaluate Eq. (37a), we take $g_a'(\nu_1,\gamma_1)\cong g_a''(\nu_1,\gamma_1)\cong g_a$; in this case the I.B. momentum spectrum has the allowed K-capture shape $\sim(1-x_a)^2x_a$ and the I.B. photon-neutrino angular correlation function is

$$1+\left(\frac{(g_V{}^{(\mu)})^2-(g_A{}^{(\mu)})^2}{(g_V{}^{(\mu)})^2+3(g_A{}^{(\mu)})^2}\right)\nu_1\cdot\gamma_1. \tag{38}$$

Equation (38) is rigorous for radiative muon capture by protons and has been derived directly for this case by Huang, Yang, and Lee[27] (H.Y. and L.). As a second approximation, we can use the techniques of Eqs. (7a) to (11b) and obtain for the heavier nuclei, $Z>6$, $A>12$

$$g_a'(\nu_1,\gamma_1)\cong g_a''(\nu_1,\gamma_1)\cong\left(\frac{A-Z}{2A}\right)\left(\frac{d}{r_0}\right)^3\times\left[1-\frac{1}{10}(\langle\nu'\rangle_a r_0\nu_1+\langle\gamma\rangle_a r_0\gamma_1)^2\left(\frac{d}{r_0}\right)^2+\cdots\right] \tag{39a}$$

entirely analogous to the result [Eqs. (6a), (11a), and (11b)]

$$g_a\cong\left(\frac{A-Z}{2A}\right)\delta_a\cong\left(\frac{A-Z}{2A}\right)\left(\frac{d}{r_0}\right)^3\times\left[1-\frac{1}{10}(\nu_a r_0)^2\left(\frac{d}{r_0}\right)^2+\cdots\right]. \tag{39b}$$

Equations (37a), (39a), and (39b) yield

$$\frac{\Lambda_{\text{rad}}{}^{(\mu)}(a;\langle\gamma\rangle_a)dx_a d\gamma_1}{\Lambda^{(\mu)}(a)}\cong\left[\left(\frac{1}{4\pi^2}\frac{1}{137}\right)\left(\frac{\langle\gamma\rangle_a{}^{\max}}{m_\mu}\right)^2(1-x_a)^2x_a dx_a d\gamma_1\right]\times\left\{1-(1-x_a)x_a\left\{\frac{1}{5}+\frac{1}{15}\left[\frac{(g_A{}^{(\mu)})^2-(g_V{}^{(\mu)})^2}{(g_V{}^{(\mu)})^2+3(g_A{}^{(\mu)})^2}\right]\right\}\right\}\times\left[\frac{\left(\frac{A-Z}{2A}\right)\delta_a}{1-\left(\frac{A-Z}{2A}\right)\delta_a}\right]\left[\left(\frac{d}{r_0}\right)^2(\langle\gamma\rangle_a{}^{\max}r_0)^2\right] \tag{40}$$

which shows that the correction factor to the allowed K-capture shape for the I.B. momentum spectrum involves [as does the I.B. photon-neutrino angular correlation, Eq. (38)] the quantity $(g_A{}^{(\mu)}/g_V{}^{(\mu)})^2$. Taking

$$(g_A{}^{(\mu)}/g_V{}^{(\mu)})^2\cong(g_A{}^{(\beta)}/g_V{}^{(\beta)})^2=(1.21)^2\,{}^{11};$$

$$(d/r_0)^2=(1.47)^2\cong(\delta_a)^{\frac{2}{3}},$$

$$\delta_a=3.0\ \text{[Eqs. (11b) and (14)]};$$

$$\langle\gamma\rangle_a{}^{\max}=\langle\nu\rangle_a\cong\tfrac{3}{4}m_\mu,$$

$$r_0\cong\frac{0.67}{m_\mu}\ \text{[Eq. (11b) } et\ seq.\text{]}$$

and $(A-Z)/2A\cong\frac{1}{4}$, Eq. (40) gives

$$\frac{\Lambda_{\text{rad}}{}^{(\mu)}(a;\langle\gamma\rangle_a)dx_a}{\Lambda^{(\mu)}(a)}\cong\left[\left(\frac{1}{\pi}\frac{1}{137}\right)\left(\frac{3}{4}\right)^2(1-x_a)^2x_a dx_a\right]\times[1-0.35(1-x_a)x_a] \tag{41}$$

so that the ratio of the total radiative capture rate,

$$\Lambda_{\text{rad}}{}^{(\mu)}(a)=\int_0^1\Lambda_{\text{rad}}{}^{(\mu)}(a,\langle\gamma\rangle_a)dx_a,$$

to the total nonradiative capture rate is

$$\frac{\Lambda_{\text{rad}}{}^{(\mu)}(a)}{\Lambda^{(\mu)}(a)}\cong\left[\left(\frac{1}{12\pi}\frac{1}{137}\right)\left(\frac{3}{4}\right)^2\right]\left[1-\frac{0.35}{5}\right]=1\times10^{-4}. \tag{42}$$

Equations (40), (41), and (42) hold also for radiative muon capture by a proton upon omission of the term $(\langle\gamma\rangle_a{}^{\max}/m_\mu)^2=(\frac{3}{4})^2$ and of the term $\sim(1-x_a)x_a$, $-0.35/5$—this has been shown directly in the case of Eq. (42) by H.Y. and L.[27] and may be seen from Eq. (37a) since for a proton $\langle\gamma\rangle_a{}^{\max}/m_\mu=0.94\cong1$ and $g_a'(\nu_1,\gamma_1)=0$, $g_a''(\nu_1,\gamma_1)=0$ (as $Z=A$).

Cantwell[25] has given, on the basis of the evaluation of Eqs. (37a) to (37d) with a simple variational trial form for

$$a|\rangle\cong\Phi_a(\mathbf{r}_1,\mathbf{r}_2,\cdots)X_a(\sigma_1{}^{(3)},\sigma_2{}^{(3)},\cdots;\tau_1{}^{(3)},\tau_2{}^{(3)},\cdots)$$

[Eqs. (16) *et seq.*], an expression for

$$\frac{\Lambda_{\text{rad}}{}^{(\mu)}(a,\langle\gamma\rangle_a)dx_a}{\Lambda^{(\mu)}(a)}$$

[27] Huang, Yang, and Lee, Phys. Rev. 108, 1348 (1957).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

in the case $\mu^-+He_2^4 \rightarrow \gamma+\nu+H_1^4$. This expression corresponds to a correction factor to the allowed K-capture shape $\cong[1-1.8(1-x_a)x_a]$ (for $(g_A^{(\mu)}/g_V^{(\mu)})^2 \cong(g_A^{(\beta)}/g_V^{(\beta)})^2=(1.21)^2$ [11]) rather than

$$\cong[1-0.35(1-x_a)x_a]$$

as predicted by Eq. (41) for the heavier nuclei. We should also mention that all of the results in Eqs. (36a) to (42) refer, in the case $J_a\neq 0$, to the appropriate averages [as in Eq. (26)] of the various radiative and nonradiative muon capture rates over the two $(J_a+\frac{1}{2}, J_a-\frac{1}{2})$ hyperfine states of the parent mu-mesic atom.

The I.B. momentum spectrum of Eqs. (40) or (41), is susceptible to observational test—a plot of experimental values of $\Lambda_{rad}^{(\mu)}(a;\langle\gamma\rangle_a)/[(1-x_a)^2x_a]$ *vs* the corresponding $(1-x_a)x_a$ should, if our approximations are sufficiently accurate, lie on a straight line with a slope whose value yields $(g_A^{(\mu)}/g_V^{(\mu)})^2$ (subject to the uncertainty in the numerical value of δ_a). This I.B. spectrum peaks at $x_a\cong\frac{1}{3}$ or $\langle\gamma\rangle_a\cong 27$ Mev and still has very appreciable values at $\langle\gamma\rangle_a\geqq 55$ Mev, i.e., at I.B. photon energies high compared to the energies of any numerous background photons. This fact should permit detection of the muon radiative capture in spite of the relative rarity of the phenomenon.

In concluding the present section it is important to emphasize the interest of a calculation of radiative muon capture using an effective Hamiltonian which includes all nucleon recoil corrections (terms $\sim\nu/m_p$) and in particular contains the "induced" pseudoscalar and the "conserved vector current" anomalous magnetic moment contributions. Such a calculation is now being carried out by Bernstein[28] and is expected to exhibit additional terms in the correction factor to the allowed K-capture shape for the I.B. momentum spectrum.

7. PARITY NONCONSERVATION EFFECTS

The anticipated presence of parity nonconservation effects in maximum amounts in the muon capture process is incorporated into the effective Hamiltonian of Eq. (1a) through the assumption that the emitted neutrino carries unit negative helicity—this is expressed mathematically by the (two-component neutrino coupling type) spin projection operator $(1-\boldsymbol{\sigma}\cdot\boldsymbol{\nu}_1)/\sqrt{2}$ in the $H_{eff}^{(\mu)}$ of Eq. (1a). No direct experimental test of this assumption is as yet available for muon capture but it has now been established that neutrinos emitted in the analogous processes of electron orbital capture and positron beta decay do possess a helicity $=-1$.[11] It would clearly be of great interest to observe parity nonconservation effects in muon capture and the present section is devoted to a discussion of four phenomena in which pseudoscalar quantities are to be measured—cf. Eqs. (44), (53), (57), and (59).

(a) Angular Distribution of Recoil Nuclei in Capture of Polarized Muons

Experimental evidence is now available that negative muons still retain an appreciable fraction of their spin polarization at the instant of decay or capture from the lowest Bohr orbit of the parent mu-mesic atom. This evidence is based on the observation of an anisotropic angular distribution of the decay electrons relative to a unit vector, $\mathbf{s}_{\mu;1}$, in the direction of the muon spin,

$$1-P_\mu\tfrac{1}{3}(\mathbf{s}_{\mu;1}\cdot\mathbf{p}_{el;1}), \tag{43a}$$

$$(\mathbf{s}_{\mu;1}\cdot\mathbf{p}_{neg.\ muon;1})=+1,^{29} \tag{43b}$$

and corresponds to a residual muon polarization at the instant of decay or capture, P_μ, of about 15 to 20% for the case of various spin zero parent nuclei.[30] The angular distribution of recoil daughter nuclei in a particular state, the recoils being formed in muon capture by spin zero parent nuclei, is then also expected to exhibit an anisotropy relative to $\mathbf{s}_{\mu;1}$, *viz.*,

$$1+P_\mu\alpha(a\rightarrow b)(\mathbf{s}_{\mu;1}\cdot\mathbf{p}_{rec;1}), \tag{44}$$

where $\mathbf{p}_{rec;1}$ is a unit vector in the direction of the recoiling nucleus and the anisotropy coefficient, $\alpha(a\rightarrow b)$, is a quantity involving the muon capture nuclear matrix elements of Eq. (2b). To avoid complications associated with the "hyperfine" effect (Sec. 5) we confine our discussion until further notice to the case of zero spin parent nuclei—this case is in addition distinguished by a lack of hyperfine-structure induced muon depolarization which, for example if $J_a=\frac{1}{2}$, cuts down the otherwise effective value of P_μ by a factor of at least 2 [see Eqs. (52b) to (52d)].

Calculation of $\alpha(a\rightarrow b)$ on the basis of the $H_{eff}^{(\mu)}$ of Eq. (1a) yields

$$\alpha(a\rightarrow b)\cong\frac{\sum_{M_b,M_a}\{(G_V^{(\mu)})^2|\langle b|\exp|a\rangle|^2+\frac{1}{3}(-(G_A^{(\mu)})^2+(G_P^{(\mu)})^2-2G_A^{(\mu)}G_P^{(\mu)})|\langle b|(\exp)\boldsymbol{\sigma}|a\rangle|^2\}}{\sum_{M_b,M_a}\{(G_V^{(\mu)})^2|\langle b|\exp|a\rangle|^2+\frac{1}{3}(3(G_A^{(\mu)})^2+(G_P^{(\mu)})^2-2G_A^{(\mu)}G_P^{(\mu)})|\langle b|(\exp)\boldsymbol{\sigma}|a\rangle|^2\}} \tag{45}$$

[28] J. Bernstein (to be published).
[29] Culligan, Frank, and Holt, Conference on Weak Interactions, Gatlinburg, Tennessee (1958, to be published); Macq, Crowe, and Haddock, Phys. Rev. **112**, 2061 (1958).
[30] Garwin, Lederman, and Weinrich, Phys. Rev. **105**, 1415 (1957); Ignatenko, Yegorov, Khalupa, and Chulthem, J. Exptl. Theoret. Phys. (U.S.S.R.) **35**, 5 (1958).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

subject to the assumption that the wave functions $|a; J_a=0^+\rangle$, $|b; J_i\rangle$ are such that

$$\sum_{M_b,M_a} |\langle b|(\exp)\boldsymbol{\sigma}\cdot\boldsymbol{\nu}_1|a\rangle|^2 \cong \tfrac{1}{3}\sum_{M_b,M_a}|\langle b|(\exp)\boldsymbol{\sigma}|a\rangle|^2.\ ^{1}$$

Formulas equivalent to that in Eq. (45), but without inclusion of the anomalous magnetic moment terms in the corresponding $H_{\rm eff}{}^{(\mu)}$ and sometimes without the term in $G_P{}^{(\mu)}$, have been given by Ioffe,[31] H. Y. and L.,[27] Shapiro, Dolinsky, and Blokhintsev,[32] Wolfenstein,[33] Überall,[34] Treiman,[35] and Fulton.[36]

In a case such as the ground state ($J_a=0^+$) to ground state ($J_b=1^+$) transition in $\mu^-+C_6{}^{12}\rightarrow\nu+B_5{}^{12}$ (Sec. 4) the spin independent nuclear matrix element $\langle b|\exp|a\rangle$ [Eq. (36b)] vanishes[1] and Eq. (45) becomes [using also Eqs. (1b) and (1c) with $\nu_{ba}=0.86\,m_\mu$ [Eq. (2c)], and $(g_A{}^{(\beta)}/g_V{}^{(\beta)})=-1.21$ [11])

$$\alpha(a\rightarrow b)=\frac{-(G_A{}^{(\mu)})^2+(G_P{}^{(\mu)})^2-2G_A{}^{(\mu)}G_P{}^{(\mu)}}{3(G_A{}^{(\mu)})^2+(G_P{}^{(\mu)})^2-2G_A{}^{(\mu)}G_P{}^{(\mu)}}=-0.73. \tag{46}$$

Thus an accurate measurement of the angular distribution of the ground state $B_5{}^{12}$ recoils, P_μ being determined by a parallel measurement of the muon decay electron angular distribution [Eq. (43a)], would yield information about the ratio $G_P{}^{(\mu)}/G_A{}^{(\mu)}$ and hence about the ratio $g_P{}^{(\mu)}/g_A{}^{(\beta)}$ [Eq. (1c) *et seq.*]. The corresponding $\alpha(a\rightarrow b)$ [Eq. (46)] is, fortunately, quite sensitive to the exact value of $g_P{}^{(\mu)}/g_A{}^{\beta}$ and to the omission or inclusion of the terms $\sim(\mu_p-\mu_n)$ being, for example, -0.33 if $g_P{}^{(\mu)}/g_A{}^{(\beta)}=0$ and if the terms $\sim(\mu_p-\mu_n)$ are absent.

We now apply the closure approximation to find the anisotropy coefficient, $\alpha(a)$,

$$\alpha(a)\equiv\frac{\sum_b\{(G_V{}^{(\mu)})^2|\langle b|\exp|a\rangle|^2+\frac{1}{3}(-(G_A{}^{(\mu)})^2+(G_P{}^{(\mu)})^2-2G_A{}^{(\mu)}G_P{}^{(\mu)})|\langle b|(\exp)\boldsymbol{\sigma}|a\rangle|^2\}}{\sum_b\{(G_V{}^{(\mu)})^2|\langle b|\exp|a\rangle|^2+\frac{1}{3}(3(G_A{}^{(\mu)})^2+(G_P{}^{(\mu)})^2-2G_A{}^{(\mu)}G_P{}^{(\mu)})|\langle b|(\exp)\boldsymbol{\sigma}|a\rangle|^2\}}$$

$$=\frac{(G_V{}^{(\mu)})^2\langle a|[\exp]_a{}^+[\exp]_a|a\rangle+\frac{1}{3}(-(G_A{}^{(\mu)})^2+(G_P{}^{(\mu)})^2-2G_A{}^{(\mu)}G_P{}^{(\mu)})\langle a|[(\exp)\boldsymbol{\sigma}]_a{}^+\cdot[(\exp)\boldsymbol{\sigma}]_a|a\rangle,}{(G_V{}^{(\mu)})^2\langle a|[\exp]_a{}^+[\exp]_a|a\rangle+\frac{1}{3}(3(G_A{}^{(\mu)})^2+(G_P{}^{(\mu)})^2-2G_A{}^{(\mu)}G_P{}^{(\mu)})\langle a|[\exp)\boldsymbol{\sigma}]_a{}^+\cdot[(\exp)\boldsymbol{\sigma}]_a|a\rangle}, \tag{47a}$$

where

$$[\exp]_a\equiv\textstyle\sum_i\tau_i{}^{(-)}\exp(-i\langle\nu\rangle_a\boldsymbol{\nu}_1\cdot\mathbf{r}_i)\varphi(r_i);\quad [(\exp)\boldsymbol{\sigma}]_a\equiv\sum_i\tau_i{}^{(-)}\exp(-i\langle\nu\rangle_a\boldsymbol{\nu}_1\cdot\mathbf{r}_i)\varphi(r_i)\boldsymbol{\sigma}_i \tag{47b}$$

which enters into the recoil nucleus angular distribution, $1+P_\mu\alpha(a)(\mathbf{s}_{\mu;1}\cdot\mathbf{p}_{\rm rec;1})$, appropriate to the total muon capture rate by the parent nucleus. Employing the techniques of Eqs. (4a) to (11b) then gives, for the heavier nuclei, $Z>6$, $A>12$ {using also Eqs. (1b) and (1c) with $\langle\nu\rangle_a=0.75\,m_\mu$ [Eq. (3d)] and $g_A{}^{(\beta)}/g_V{}^{(\beta)}=-1.21$ [11]},

$$\alpha(a)\cong\frac{(G_V{}^{(\mu)})^2-(G_A{}^{(\mu)})^2+(G_P{}^{(\mu)})^2-2G_A{}^{(\mu)}G_P{}^{(\mu)}}{(G_V{}^{(\mu)})^2+3(G_A{}^{(\mu)})^2+(G_P{}^{(\mu)})^2-2G_A{}^{(\mu)}G_P{}^{(\mu)}}$$

$$=-0.39. \tag{48}$$

Further, if the daughter nucleus is unbound even in its ground state, e.g.,

$$\mu^-+Mg_{12}{}^{26}\rightarrow\nu+\{Na_{11}{}^{26}\rightarrow Na_{11}{}^{25}+n_0{}^1\}, \tag{49}$$

it is not unreasonable to suppose that in the great majority of energetically accessible final states a single neutron carries off most of the available recoil momentum. Under such circumstances the angular distribution of the recoiling neutrons is, approximately,

$$1+P_\mu\alpha(a)(\mathbf{s}_{\mu;1}\cdot\mathbf{p}_{\rm neu;1}), \tag{50}$$

where $\alpha(a)$ is as in Eq. (48) and $\mathbf{p}_{\rm neu;1}$ is a unit vector in the recoil neutron direction. Überall[34] has considered, on the basis of a Fermi gas model, the effective interaction of the recoiling neutron with the remaining nucleons and has concluded that this effective interaction is probably not sufficiently strong to distort appreciably the angular distribution in Eq. (50).

We now consider recoil anisotropy for the case of muon capture by parent nuclei with nonzero spin. Here the anisotropy coefficients are different in the two different hyperfine states of the parent mu-mesic atom, i.e., one must distinguish between $\alpha(J_a+\frac{1}{2};a)$ and $\alpha(J_a-\frac{1}{2};a)$. As a general rule it is obvious that $\alpha(J_a-\frac{1}{2};a)=0$ for $J_a=\frac{1}{2}$ since the corresponding parent atom hyperfine state is then a singlet and so is spatially spherical. As a particular illustration we may give the formulas for the case of the hydrogen mu-mesic atom $[p\mu^-]$. Here, by the remark just made,

$$\alpha(\tfrac{1}{2}-\tfrac{1}{2};H_1{}^1)=0, \tag{51a}$$

[31] B. L. Ioffe, J. Exptl. Theoret. Phys. (U.S.S.R.) 33, 308 (1957).
[32] Shapiro, Dolinsky, and Blokhintsev, Nuclear Phys. 4, 273 (1957).
[33] L. Wolfenstein, Nuovo cimento 7, 706 (1958).
[34] H. Überall, Nuovo cimento 6, 533 (1957).
[35] S. B. Treiman, Phys. Rev. 110, 448 (1958).
[36] T. Fulton, Nuclear Phys. 6, 319 (1958).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

and, as shown in a calculation by Bernstein and Primakoff,[37]

$$\alpha(\tfrac{1}{2}+\tfrac{1}{2};\,H_1{}^1)=\frac{\tfrac{1}{2}[(G_V{}^{(\mu)}+G_A{}^{(\mu)})^2+(G_P{}^{(\mu)})^2-2(G_V{}^{(\mu)}+G_A{}^{(\mu)})G_P{}^{(\mu)}]}{\tfrac{3}{4}[(G_V{}^{(\mu)}+G_A{}^{(\mu)})^2+(G_P{}^{(\mu)})^2-\tfrac{2}{3}(G_V{}^{(\mu)}+G_A{}^{(\mu)})G_P{}^{(\mu)}]}. \tag{51b}$$

Thus at low $(H_1{}^1)_2$ molecule number densities, where there is no appreciable collisional hyperfine triplet to singlet conversion (Sec. 5), the angular distribution of the recoil neutrons is:

$$\frac{\tfrac{3}{4}\Lambda^{(\mu)}(\tfrac{1}{2}+\tfrac{1}{2};\,H_1{}^1)[1+P_\mu\alpha(\tfrac{1}{2}+\tfrac{1}{2};\,H_1{}^1)(\mathbf{s}_{\mu;1}\cdot\mathbf{p}_{\text{neu};1})]+\tfrac{1}{4}\Lambda^{(\mu)}(\tfrac{1}{2}-\tfrac{1}{2};\,H_1{}^1)}{\Lambda^{(\mu)}(H_1{}^1)} \tag{52a}$$

or {using also Eqs. (33a), (33b), (27b), (1b), and (1c) (with $\langle\nu\rangle_a=0.94\,m_\mu$ [Eq. (2c)], and $(g_A{}^{(\beta)}/g_V{}^{(\beta)})=-1.21$ [11])},

$$1+\frac{P_\mu}{2}\frac{[(G_V{}^{(\mu)}+G_A{}^{(\mu)})^2+(G_P{}^{(\mu)})^2-2(G_V{}^{(\mu)}+G_A{}^{(\mu)})G_P{}^{(\mu)}]}{[(G_V{}^{(\mu)})^2+3(G_A{}^{(\mu)})^2+(G_P{}^{(\mu)})^2-2G_A{}^{(\mu)}G_P{}^{(\mu)}]}(\mathbf{s}_{\mu;1}\cdot\mathbf{p}_{\text{neu};1}) \tag{52b}$$

$$=1+\frac{P_\mu}{2}[0.01](\mathbf{s}_{\mu;1}\mathbf{p}_{\text{neu};1}), \tag{52c}$$

while the corresponding muon decay electron angular distribution becomes

$$1-\frac{P_\mu}{2}\tfrac{1}{3}(\mathbf{s}_{\mu;1}\cdot\mathbf{p}_{\text{el};1}). \tag{52d}$$

P_μ must be interpreted in Eqs. (52a) to (52d) as what the residual muon polarization would be, at the instant of decay or capture, if the parent proton did not create any hyperfine-structure induced depolarization in the muon $1s_{\frac{1}{2}}$ orbit—it is seen that the numerical values of $G_V{}^{(\mu)}$, $G_A{}^{(\mu)}$, $G_P{}^{(\mu)}$ are such that the recoil neutron distribution is practically isotropic; this is expected since $[\Lambda^{(\mu)}(\frac{1}{2}+\frac{1}{2};\,H_1{}^1)/\Lambda^{(\mu)}(\frac{1}{2}-\frac{1}{2};\,H_1{}^1)]\ll 1$ [Eqs. (33a) and (33b)]. At higher $(H_1{}^1)_2$ molecule densities all the $[p\mu^-]$ are collisionally converted into the hyperfine singlet state (Sec. 5), so that as already noted above and as first pointed out by Gershtein and Zeldovich,[22] the recoil neutron distribution is here certainly expected to be isotropic.

(b) Angular Distribution of Photons in Radiative Capture of Polarized Muons

The internal bremsstrahlung (I.B.) photons emitted in the radiative capture of polarized muons are also expected to be characterized by an anisotropic angular distribution relative to $\mathbf{s}_{\mu;1}$; this distribution is, for the case of zero spin parent nuclei with the daughter nuclei formed in a particular state:

$$1+P_\mu\beta(a\to b;\,\gamma_{ba})(\mathbf{s}_{\mu;1}\cdot\boldsymbol{\gamma}_1), \tag{53}$$

where $\boldsymbol{\gamma}_{ba}=\gamma_{ba}\boldsymbol{\gamma}_1$ is the I.B. photon momentum vector, $\beta(a\to b;\,\gamma_{ba})$ the corresponding anisotropy coefficient, and, for the reasons mentioned in Sec. 7(a), the discussion is again confined to the case of zero spin parent nuclei.

The anisotropy coefficient $\beta(a\to b;\,\gamma_{ba})$ of Eq. (53) may be directly calculated or may be obtained on the basis of a theorem of Cutkosky[38] which shows that this $\beta(a\to b;\,\gamma_{ba})$ is numerically equal to the helicity of a *massless* positron emitted in the beta-decay process: $|a;Z,A\rangle\to|b;Z-1,A\rangle+e^+\!+\nu$ ($Z/137\ll 1$; beta-decay process allowed or forbidden). Now in a beta-decay theory described by an effective Hamiltonian of the type of Eq. (1a) with $G_V{}^{(\mu)}\to g_V{}^{(\beta)}$, $G_A{}^{(\mu)}\to g_A{}^{(\beta)}$, $G_P{}^{(\mu)}\to 0$ the helicity of such a massless positron is $+1$ —thus in a muon radiative capture theory with an $H_{\text{eff}}{}^{(\mu)}$ characterized by $G_V{}^{(\mu)}=g_V{}^{(\mu)}$, $G_A{}^{(\mu)}=g_A{}^{(\mu)}$, $G_P{}^{(\mu)}=0$, i.e., with an $H_{\text{eff}}{}^{(\mu)}$ in which all nucleon recoil effects (terms $\sim\nu/m_p$) are omitted, one has

$$\beta(a\to b;\,\gamma_{ba})=1. \tag{54}$$

Thus, summing over all the energetically accessible states of the daughter nucleus, we obtain

$$\beta(a;\,\langle\boldsymbol{\gamma}\rangle_a)=\frac{\sum_b\beta(a\to b;\,\gamma_{ba})\Lambda_{\text{rad}}{}^{(\mu)}(a\to b;\,\gamma_{ba}}{\sum_b\Lambda_{\text{rad}}{}^{(\mu)}(a\to b;\,\gamma_{ba})}=1, \tag{55}$$

which last pair of equations have also been explicitly derived by H.Y. and L.[27] and by Bernstein.[28]

There is now very considerable interest in a calculation of $\beta(a\to b;\,\gamma_{ba})$, $\beta(a;\,\langle\boldsymbol{\gamma}\rangle_a)$, where one includes nucleon recoil terms $\sim\nu/m_p$ into the appropriate effective Hamiltonian, i.e., includes in particular the induced pseudoscalar and the conserved vector current anomalous magnetic moment contributions—such a calculation is being carried out by Bernstein.[28] The effect of the pseudoscalar term can be foreseen quali-

[37] J. Bernstein and H. Primakoff (to be published).

[38] R. E. Cutkosky, Phys. Rev. **107**, 330 (1957).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

tatively on the basis of Cutkosky's theorem[38] since it is known that with an effective V, A, P beta-decay interaction the helicity of a massless positron is less than unity; hence in a theory with $g_P{}^{(\mu)}\neq 0$ the calculated values of $\beta(a\rightarrow b;\gamma_{ba})$, $\beta(a;\langle\gamma\rangle_a)$ will also be less than unity. Thus any measurement of the I.B. photon angular distribution sufficiently accurate to fix a reliable value for $(1-\beta(a;\langle\gamma\rangle_a))$ will, upon comparison with Bernstein's theoretical expression[28] for $\beta(a;\langle\gamma\rangle_a)$, yield an "experimental" value of $g_P{}^{(\mu)}$ which can be compared with the Goldberger-Treiman[3]-Wolfenstein[4] theoretical value of $8g_A{}^{(\beta)}$ [Eq. (1c)].

(c) Polarization of Recoil Nuclei in Muon Capture

The recoil daughter nuclei formed in muon capture are in general polarized even when the polarization of the muon itself at the instant of capture is negligible. This effect has been discussed in a fairly general way by Treiman[35] and by Fulton,[36] and for the particular case of the hydrogen mu-mesic atom by Gershtein and Zeldovich[22] and by Bernstein and Primakoff.[37] For $[p\mu^-]$ in the hyperfine singlet state Gershtein and Zeldovich[22] point out that conservation of angular momentum and the assumption that the neutrino has unit negative helicity ensures, independent of the magnitudes of the various coupling constants in $H_{\rm eff}{}^{(\mu)}$, that the helicity of the recil neutron is $=-1$. For $[p\mu^-]$ in the hyperfine triplet state an explicit calculation by Bernstein and Primakoff[37] shows that the recoil neutron helicity is

$$-\frac{1}{3}\frac{(G_V{}^{(\mu)}+G_A{}^{(\mu)})^2+(G_P{}^{(\mu)})^2-6(G_V{}^{(\mu)}+G_A{}^{(\mu)})G_P{}^{(\mu)}}{(G_V{}^{(\mu)}+G_A{}^{(\mu)})^2+(G_P{}^{(\mu)})^2-\frac{2}{3}(G_V{}^{(\mu)}+G_A{}^{(\mu)})G_P{}^{(\mu)}} \quad (56a)$$

so that at low $(H_1{}^1)_2$ molecule number densities the over-all recoil neutron helicity is {using also Eqs. (33a), (33b), (27b), (1b), and (1c) (with $\langle\nu\rangle_a=0.94\,m_\mu$ [Eq. (2c)] and $(g_A{}^{(\beta)}/g_V{}^{(\beta)})=-1.21$ [11])}

$$\frac{3}{4}\frac{\Lambda^{(\mu)}(\frac{1}{2}+\frac{1}{2};H_1{}^1)}{\Lambda^{(\mu)}(H_1{}^1)}\left[-\frac{1}{3}\frac{(G_V{}^{(\mu)}+G_A{}^{(\mu)})^2+(G_P{}^{(\mu)})^2-6(G_V{}^{(\mu)}+G_A{}^{(\mu)})G_P{}^{(\mu)}}{(G_V{}^{(\mu)}+G_A{}^{(\mu)})^2+(G_P{}^{(\mu)})^2-\frac{2}{3}(G_V{}^{(\mu)}+G_A{}^{(\mu)})G_P{}^{(\mu)}}\right]+\frac{1}{4}\frac{\Lambda^{(\mu)}(\frac{1}{2}-\frac{1}{2};H_1{}^1)}{\Lambda^{(\mu)}(H_1{}^1)}[-1]$$

$$=2\frac{G_V{}^{(\mu)}G_A{}^{(\mu)}-(G_A{}^{(\mu)})^2-G_V{}^{(\mu)}G_P{}^{(\mu)}}{(G_V{}^{(\mu)})^2+3(G_A{}^{(\mu)})^2+(G_P{}^{(\mu)})^2-2G_A{}^{(\mu)}G_P{}^{(\mu)}}=-0.99. \quad (56b)$$

This is very close to -1 as is indeed expected from the fact that $[\Lambda^{(\mu)}(\frac{1}{2}+\frac{1}{2};H_1{}^1)/\Lambda^{(\mu)}(\frac{1}{2}-\frac{1}{2};H_1{}^1)]\ll 1$ [Eqs. (33a) and (33b)].

In the calculation given by Treiman[35] and by Fulton[36] the $H_{\rm eff}{}^{(\mu)}$ of Eq. (1a) with $G_V{}^{(\mu)}=g_V{}^{(\mu)}$, $G_A{}^{(\mu)}=g_A{}^{(\mu)}$, $G_P{}^{(\mu)}=0$, has been used; also, for the reasons mentioned in Sec. 7a, the discussion should be confined to the case of zero spin parent nuclei. Making the additional assumption that the wave functions $|a;J_a=0^+\rangle$, $|b;J_b\rangle$ of the parent and daughter nuclei are such that the s-wave part of $\exp(-i\nu_{ba}\cdot\mathbf{r}_i)$ predominates over the d wave, g wave, $\cdots$ parts in the nuclear matrix elements,[1] so that $J_b=1^+$, one may calculate the polarization of the recoil daughter nuclei formed in a particular state with $J_b=1^+$. Since in a $J_a=0^+\rightarrow J_b=1^+$ transition, as for example $\mu^-+C_6{}^{12}=\nu+(B_5{}^{12})_{\text{ground state}}$, the spin independent nuclear matrix element $\langle b|\exp|a\rangle$ [Eq. (36b)] vanishes,[1] the spin dependent nuclear matrix element $\langle b|(\exp)\boldsymbol{\sigma}|a\rangle$ [Eq. (36b)] necessarily cancels out of the expression for the recoil polarization and this becomes

$$\langle\mathbf{J}_{b;1}(\mathbf{p}_{\text{rec};1})\rangle=-\frac{2}{3}\frac{-\mathbf{p}_{\text{rec};1}+P_\mu\mathbf{s}_{\mu;1}}{1-\frac{1}{3}P_\mu\mathbf{s}_{\mu;1}\mathbf{p}_{\text{rec};1}}. \quad (57)$$

Thus the recoil polarization is (anti) parallel to $\mathbf{p}_{\text{rec};1}$ only if the muon polarization at the instant of capture, P_μ, vanishes. One can also obtain the recoil polarization averaged over all possible directions of recoil: $\{\langle\mathbf{J}_{b;1}\rangle\}_{\text{Av}}$ —this turns out to be (Jackson, Treiman, and Wyld[39])

$$\{\langle\mathbf{J}_{b;1}\rangle\}_{\text{Av}}=\tfrac{2}{3}P_\mu\mathbf{s}_{\mu;1} \quad (58)$$

and is just what $\langle\mathbf{J}_{b;1}(\mathbf{p}_{\text{rec};1})\rangle$ would be in a theory with a parity conserving $H_{\rm eff}{}^{(\mu)}$. The quantity $\{\langle\mathbf{J}_{b;1}\rangle\}_{\text{Av}}$ has recently been measured by Love, Marder, Nadelhaft, Siegel, and Taylor[40] on the basis of the observation of the angular anisotropy, relative to $\mathbf{s}_{\mu;1}$, of the decay electrons of the daughter nucleus: $(B_5{}^{12})_{\text{ground state}}$. $\{\langle\mathbf{J}_{b;1}\rangle\}_{\text{Av}}$ was found to be positive if one identified the directions $\mathbf{s}_{\mu;1}$ and $\mathbf{p}_{\text{neg. muon};1}$ [Eq. (43b)] while its magnitude was appropriate to a reasonable amount of depolarization of the $B_5{}^{12}$ by hyperfine interaction with its atomic electrons.

(d) Polarization of Photons in Radiative Muon Capture

The internal bremsstrahlung (I.B.) photons emitted in radiative muon capture are circularly polarized independent of any residual polarization of the muon itself. Cutkosky,[38] H.Y. and L.[27] and Bernstein[28] have shown that the degree of circular polarization, $\beta'(a\rightarrow b;\gamma_{ba})$ of any I.B. photon with momentum γ_{ba}—$\beta'(a\rightarrow b;\gamma_{ba})=\pm 1$ for complete right-hand, left-hand

[39] Jackson, Treiman, and Wyld, Phys. Rev. **107**, 327 (1957).
[40] Love, Marder, Nadelhaft, Siegel, and Taylor, Phys. Rev. Letters **2**, 107 (1959).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

circular polarization—is numerically equal to the parameter $\beta(a \to b; \gamma_{ba})$ which is the corresponding anisotropy coefficient of the I.B. photon momentum directional distribution [Eq. (53)]. The conclusions drawn in Sec. (7b) about $\beta(a \to b; \gamma_{ba})$, $\beta(a; \langle\gamma\rangle_a)$ therefore can be applied to the corresponding quantities $\beta'(a \to b; \gamma_{ba})$,

$$\beta'(a; \langle\gamma\rangle_a) = \frac{\sum_b \beta'(a \to b; \gamma_{ba}) \Lambda_{\text{rad}}{}^{(\mu)}(a \to b; \gamma_{ba})}{\sum_b \Lambda_{\text{rad}}{}^{(\mu)}(a \to b; \gamma_{ba})} \quad (59)$$

so that in particular one expects deviations from complete right circular polarization of the I.B. photons only to the extent that the induced pseudoscalar interaction is present.

ACKNOWLEDGMENT

I should like to thank Professor V. L. Telegdi for many interesting and informative discussions on the subject of muon capture and for the communication of experimental data before publication.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

Strange Particle Decay Processes and the Fermi Interaction*

R. H. DALITZ

Enrico Fermi Institute for Nuclear Studies, University of Chicago, Chicago, Illinois

1. INTRODUCTION

IT recently has become apparent that the processes of beta decay, muon decay, and nuclear capture of μ^- mesons may each be well described in terms of a four-fermion coupling of the form $V-A$, each of these couplings having a strength $gm_\pi^2 \sim 10^{-7}$. Feynman and Gell-Mann[1] have suggested that these interactions may be parts of a more general weak interaction expressed in terms of a weak interaction current J_μ, this general interaction having the form

$$H_{wk}=J_\mu^* J_\mu, \tag{1.1}$$

where the current J_μ consists of the sum

$$J_\mu=g^{\frac{1}{2}}\bar{\nu}\gamma_\mu(1+\gamma_5)\mu^+ + g^{\frac{1}{2}}\bar{\nu}\gamma_\mu(1+\gamma_5)e^+ + g^{\frac{1}{2}}\bar{n}\gamma_\mu(1+\gamma_5)p+\cdots. \tag{1.2}$$

These authors, and earlier Gershtein and Zeldovich,[2] have suggested further that the terms of J_μ which involve strongly interacting particles, but allow no change of strangeness, may obey a conservation principle.

As a result of these successes of the four-fermion interaction, attention is confined here to a survey of the possibility of accounting for strange particle decays by the addition of further terms to the current J_μ, terms which have the same form as those of Eq. (1.2) but which do not conserve strangeness. There are many possibilities for additional terms of this kind, for example $\bar{\Lambda}\gamma_\mu(1+\gamma_5)p$, $\bar{\Sigma}^-\gamma_\mu(1+\gamma_5)n$, $\bar{\Xi}^-\gamma_\mu(1+\gamma_5)\Lambda$, etc. The coupling of these terms to those of (1.2) which involve strongly interacting particles and no change in strangeness then lead directly to the pionic decay modes which are observed for the hyperons; their coupling to the leptonic terms of (1.2) leads to leptonic decay modes for the hyperons. The decay of K mesons is then interpreted as due to their coupling to virtual hyperon-nucleon pairs which allow weak decays for the K mesons through the four-fermion couplings of the expression (1.2), for example

$$K^+ \xrightarrow[\text{strong}]{} p+\bar{\Lambda}\begin{cases}\xrightarrow[\text{weak}]{} \mu^+ + \bar{\nu}, & (1.3a)\\ \xrightarrow[\text{weak}]{} p+\bar{n} \xrightarrow[\text{strong}]{} \pi^+ + \pi^- + \pi^+. & (1.3b)\end{cases}$$

The possibility of a qualitative account of the decay modes of hyperons and K particles on the basis of such four-fermion couplings has been known for some time. It was first put forward by Dallaporta[3] and by Gell-Mann,[4] and has been discussed in some detail by Gell-Mann and Rosenfeld[5] in their recent review article.

At this stage it is necessary to inquire to what extent this scheme offers the possibility of accounting quantitatively[6] for the branching ratios and detailed characteristics of these decay modes. However, such a complete program is not carried through here. Our present purpose is simply to discuss the degree of agreement between the data and those expectations from this model which do not depend on detailed theoretical calculations of its consequences.

2. PIONIC DECAY MODES FOR STRANGE PARTICLES

First *the* Λ *decay modes*,

$$\Lambda \to \begin{cases} p+\pi^-, & (2.1a)\\ n+\pi^0, & (2.1b)\end{cases}$$

are considered. These modes are regarded as the result of weak interactions connecting $(\bar{\Lambda}p)$ and $(\bar{n}p)$. With a parity nonconserving weak interaction, each of these modes requires two parameters (s_-,p_-) and (s_0,p_0), which denote the amplitudes for emission of s- and p-wave pions in the π^-- and π^0-decay processes (2.1), respectively. Assuming time-reversal invariance to be valid for both strong and weak interactions, the phases of these amplitudes arise only from the scattering in the final pion-nucleon state, as first pointed out by Takeda.[7] Since the relevant pion-nucleon phase shifts are known to be small at the energy of Λ decay, the parameters s

---

* Work done under the auspices of the U. S. Atomic Energy Commission.

[1] R. Feynman and M. Gell-Mann, Phys. Rev. 109, 193 (1958).

[2] S. Gershtein and J. Zeldovich, Soviet Phys. JETP 2, 576 (1957).

[3] N. Dallaporta, Nuovo cimento 1, 962 (1953); G. Costa and N. Dallaporta, Nuovo cimento 2, 519 (1955).

[4] M. Gell-Mann, *Proceedings of the Sixth Annual Rochester Conference on High-Energy Physics* (Interscience Publishers, Inc., New York, 1956).

[5] M. Gell-Mann and A. H. Rosenfeld, Ann. Rev. Nuclear Sci. 7, 407 (1957).

[6] The absence of evidence for the beta decay of the Λ particle appeared a stumbling block for this scheme for a long time. However this process has now been observed by Crawford, Cresti, Good, Kalbfleisch, Stevenson, and Ticho [Phys. Rev. Letters 1, 377 (1958)] and by Nordin, Orear, Reed, Rosenfeld, Solmitz, Taft, and Tripp [Phys. Rev. Letters 1, 380 (1958)]. The observed rate suggests (cf. Sec. 4) that the term of J_μ associated with this process may have an amplitude of order 0.3 relative to that for beta decay of the neutron.

[7] G. Takeda, Phys. Rev. 101, 1547 (1956).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

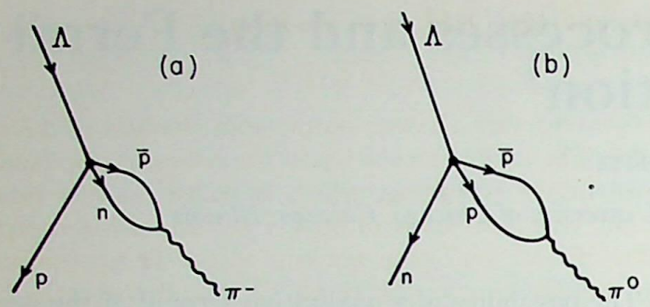


FIG. 1. The simplest graphs leading to pionic Λ decay through the Fermi interaction.

and p will be assumed real. The experimental information bearing on these parameters is at present limited to the following[8]:

$$(s_0^2+p_0^2)/(s_-^2+p_-^2)=0.59\pm0.07, \quad (2.2a)$$

$$\alpha_-=2s_-p_-/(s_-^2+p_-^2)\gtrsim 0.73\pm0.14, \quad (2.2b)$$

the first representing the branching ratio between the modes (2.1), the second representing the information available on the up-down asymmetry in the decay of polarized Λ particles. The condition (2.2b) limits p_-/s_- to lie between the limits

$$0.45\lesssim|p_-/s_-|\lesssim 2.25. \quad (2.3)$$

Calculation of the amplitudes for the processes (2.1) from the four-fermion weak couplings of the type $(\bar{\Lambda}p)(\bar{p}n)$ would involve consideration of many complicated radiative corrections arising from the strong pion couplings of these fermions, a task beyond our present ability. For the purpose of orientation, we confine attention to the simplest possible graphs leading to these decay processes, those shown in Fig. 1. For this graph the amplitude for process (2.1a) takes the form

$$\sqrt{2}C(\bar{\Lambda}\gamma_\mu(1+\gamma_5)p)\partial\phi_\pi/\partial x_\mu$$
$$=\sqrt{2}C\{(m_\Lambda-m)+(m_\Lambda+m)\boldsymbol{\sigma}\cdot\mathbf{q}/2m\}, \quad (2.4)$$

in the nonrelativistic limit appropriate to the low-energy release in Λ decay, when $\mathbf{q}$ denotes the pion momentum. This crude estimate leads to a ratio $p_-/s_-=(m_\Lambda-m)q/2m(m_\Lambda+m)=0.64$, corresponding to a value $\alpha_-=+0.9$ for the polarization coefficient in Λ decay. This value of p_-/s_- lies within the limits (2.3) from the up-down asymmetry observed[8] in polarized Λ-particle decay and corresponds to the sign recently[9] deduced for α from the polarization observed for protons resulting from the decay of unpolarized Λ particles. Such a large value of p_-/s_- is characteristic of a coupling which involves V coupling with the chirality factor $(1+\gamma_5)$; in the same approximation, S coupling (with this chirality factor) would lead to a ratio $p_-/s_-\sim q/2m=0.05$ and the large lower limit observed for α_- would then be difficult to understand in terms of a chirality factor.

Further evidence concerning this ratio p_-/s_- may be obtained from the data on the frequency of the two-body decay mode

$$_\Lambda H^4\rightarrow\pi^-+He^4, \quad (2.5)$$

relative to that for all the three-body π^--decay modes of $_\Lambda H^4$, for example π^-+p+H^3 or π^-+n+He^3. Systematic investigation by the EFINS–NU emulsion groups[10] at Chicago has established 35 examples of the two-body mode (2.5), compared with a maximum of 27 examples for other π^--mesonic modes of $_\Lambda H^4$ decay. This last figure includes six events which probably represent three-body modes of $_\Lambda H^4$ decay but which could not be uniquely established as representing $_\Lambda H^4$ decay events. From this evidence the proportion R of π^--mesonic $_\Lambda H^4$ decays which proceed through the two-body mode (2.5) is rather high, in fact,

$$R\gtrsim 0.6\pm0.1. \quad (2.6)$$

If the ground state of $_\Lambda H^4$ has spin $J=1$, the process (2.5) must involve the emission of a p-wave pion. This can take place only through the p channel of Λ decay since the initial and final nuclear systems consist predominantly of s states. According to a recent estimate,[11] the value R_1 for this case cannot exceed 0.25, reaching this value only with the upper limit (2.3) for p_-/s_-. On the other hand, with $J=0$, the pion emission can proceed through the s channel of the Λ-decay interaction and the corresponding estimate R_0 can reach a value of 0.45 with the use of the lower limit (2.3) for p_-/s_-. There is now some reason[12] to believe that these calculations may involve some overestimate of the probability of the three-body modes, and correction for this will allow correspondingly larger estimates (by as much as 20%) for R. However, the qualitative conclusion appears definite, that it is difficult to account for the value (2.6) observed for R unless $J=0$ for $_\Lambda H^4$ and $p_-/s_-<1$. With this last conclusion,

[8] D. Glaser, *Proceedings of the 1958 Annual International Conference on High Energy Physics at CERN*, p. 265; Glaser, Good, and Morrison, *Proceedings of the 1958 Annual International Conference on High Energy Physics at CERN*, p. 270.

[9] Boldt, Bridge, Caldwell, and Pal, Phys. Rev. Letters **1**, 256 (1958).

[10] Levi-Setti, Ammar, Slater, Limentani, Roberts, Schlein, and Steinberg (Nuovo cimento, to be published).

[11] R. H. Dalitz, Phys. Rev. **112**, 605 (1958).

[12] R. H. Dalitz and L. Liu (to be published). [*Note added in proof.*—Re-examination of the calculation reported in reference 11 has shown that the conclusions drawn above are too strong. With $J=0$, the values calculated for R are consistent with the experimental result (2.6), within the stated error, for values of p_-/s_- less than 1.5. With $J=1$, the calculated value of R is within two standard deviations of (2.6) for $p_-/s_->1$. Consequently, no conclusion concerning the value of p_-/s_- can be reached from these data alone until the spin of $_\Lambda H^4$ is known independently. Conversely, if a value of p_-/s_- could be obtained directly (for example, from observations on the proton polarization resulting from polarized Λ decay), this argument may allow the $_\Lambda H^4$ spin to be deduced. The evidence on the nonmesonic decay rates for Λ hypernuclei mentioned below does suggest that it is rather unlikely that p_-/s_- should exceed unity.]

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

it appears that the p_-/s_- ratio of 0.64 obtained from this crude model of Λ decay does not differ widely from the actual situation.

A similar crude calculation may be made for the mode (2.1b) on the basis of the $V-A$ interaction $(\bar{\Lambda}p)(\bar{p}n)$, with the result[13] $C(\bar{\Lambda}\gamma_\mu(1+\gamma_5)n)\partial\phi_0/\partial x_\mu$. This estimate predicts that

$$s_0/s_-=p_0/p_-=1/\sqrt{2}. \tag{2.7}$$

This is consistent with the observed ratio (2.2b) between the (π^0+n) and (π^-+p) modes of Λ decay, and it further requires that the polarization properties of these two modes should be identical, a prediction of current interest, since the ratio α_0/α_- may soon be determined from Λ particles by experiments in bubble chambers with working fluid of high density.

However, these experimental predictions would also follow directly from the $\Delta T=\frac{1}{2}$ selection rule proposed by Gell-Mann and Pais[14] for strange particle decays. With $T=0$ for the Λ particle, this would require $T=\frac{1}{2}$ for the $(\pi+N)$ system and therefore that

$$s_0/s_-=p_0/p_-=-1/\sqrt{2}. \tag{2.8}$$

It is obviously difficult to distinguish between the cases (2.7) and (2.8), since they differ only in the relative sign of the matrix elements for two different physical processes. However, this relative sign is, in principle, a physical observable and can have physical consequences, as shown in the following. At present there is no theoretical basis to favor the existence of such a $\Delta T=\frac{1}{2}$ selection rule. The proposal of this selection rule is based entirely on the empirical evidence that is discussed in the next section. It is possible to construct a four-fermion interaction which leads to this $\Delta T=\frac{1}{2}$ selection rule; in fact, for an interaction involving only the Λ hyperon, this would have the unique $(V-A)$ form

$$\{(\bar{\Lambda}p)(\bar{p}n)+(\bar{\Lambda}n)(\bar{n}n)\}. \tag{2.9}$$

However, this interaction does not fit with the viewpoint of Feynman and Gell-Mann as expressed in Eq. (1.1), since the only terms which they permit for the current J_μ involve a charge change ΔQ of one unit. This restriction appears necessary in order to forbid certain types of decay process (e.g., $\mu^+\rightarrow e^++e^-+e^+$, $K^+\rightarrow\pi^++\nu+\bar{\nu}$) whose occurrence has not been observed, and excludes terms of the type $(\bar{\nu}\nu)$, $(\bar{\Lambda}n)$ or $(\bar{n}n)$ from J_μ. The four-fermion interaction allowed by these considerations is composed of the current terms $(\bar{\Lambda}p)$ and $(\bar{p}n)$, leading to the form

$$(\bar{\Lambda}p)(\bar{p}n). \tag{2.10}$$

This interaction allows isotopic spin changes of $\Delta T=\frac{1}{2}$ and $\frac{3}{2}$. The crude calculation of the Λ-decay amplitudes given in the foregoing, which is based on this interaction, therefore leads to both $T=\frac{1}{2}$ and $T=\frac{3}{2}$ components in the final pion-nucleon state. These combine fortuitously to give the result (2.7) and the ratio 2:1 for the $(\pi^-+p)/(\pi^0+n)$ modes, in agreement with experiment. Inclusion of radiative corrections may be expected to modify the (π^0+n) and (π^-+p) amplitudes in quite different ways. On this view, the agreement between (2.2a) and the prediction of the $\Delta T=\frac{1}{2}$ rule would be regarded as fortuitous and would give no basis for expecting α_0/α_- to be unity, as would follow if the $\Delta T=\frac{1}{2}$ rule held. Observation of the polarization properties of $\Lambda\rightarrow n+\pi^0$ decay will therefore have considerable relevance for understanding of the $(\pi^-+p)/(\pi^0+n)$ ratio in Λ decay and of the basic mechanism underlying Λ decay.

The process of *nonmesonic decay of* Λ *hypernuclei* is of special interest concerning the mechanism underlying Λ decay, since the elementary four-fermion interactions (2.9) and (2.10) mentioned above actually correspond to processes of nonmesonic de-excitation of the Λ particle,[15] thus

$$\Lambda+p\rightarrow n+p \tag{2.11}$$

from interaction (2.10). The strength $\sqrt{f}$ appropriate to a current term $(\bar{\Lambda}p)$ in J_μ is not well-known empirically. The hypothesis of universality for the four-fermion weak interactions would require that f should equal g, the coupling strength appropriate to the terms $(\bar{\nu}e^+)$ or $(\bar{\nu}\mu^+)$; in fact, as shown later (cf. §4), there is reason to believe that f may be an order of magnitude smaller than g. To a sufficient approximation (of order 10%), it is sufficient to consider only the non-relativistic approximation to the $(V-A)$ interaction for (2.10), namely,

$$\begin{aligned}(fg)^{\frac{1}{2}}(\bar{\psi}_\Lambda\gamma_\mu(1+\gamma_5)\psi_p)(\bar{\psi}_p\gamma_\mu(1+\gamma_5)\psi_n)\\ \simeq(fg)^{\frac{1}{2}}\{\bar{\psi}_\Lambda\psi_p\bar{\psi}_p\psi_n-(\bar{\psi}_\Lambda\sigma_\Lambda\psi_p)\cdot(\bar{\psi}_p\sigma_N\psi_n)\}.\end{aligned} \tag{2.12}$$

This dominant term is parity conserving (conventionally, the Λ parity has been defined to be even); parity nonconserving terms are of order $v/c\sim0.25$ relative to this term. There are many radiative corrections due to the strong pion and K-meson couplings of these particles, which modify the form of the amplitude for the nonmesonic capture process (2.11) from this form (2.12). Several of the many possibilities are sketched in Fig. 2. Terms corresponding to Figs. 2(b) and 2(c) are of special importance and have been discussed in detail by Karplus and Ruderman.[16] They may be

[13] With the $V-A$ form of interaction, the interactions $(\bar{a}b)(\bar{c}d)$ and $(\bar{a}d)(\bar{b}c)$ are identical. Consequently, with $(\bar{\Lambda}n)(\bar{p}p)$ replacing $(\bar{\Lambda}p)(\bar{p}n)$, the same factor C appears here, but without the factor $\sqrt{2}$ since the emission is now of a π^0 meson instead of a π^- meson.

[14] M. Gell-Mann and A. Pais, *Proceedings of the International Conference on High Energy Physics* (Pergamon Press, London, 1955).

[15] S. B. Treiman, *Proceedings of the 1958 Annual International Conference on High Energy Physics at CERN*, p. 276.

[16] M. Ruderman and R. Karplus, Phys. Rev. **76**, 1458 (1949).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

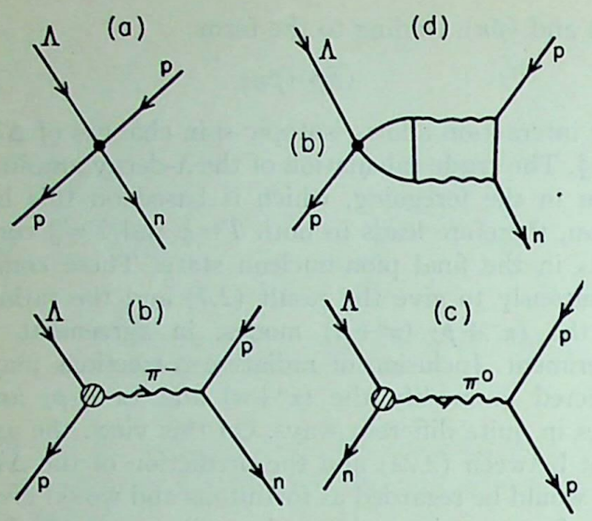


FIG. 2. The elementary four-fermion interaction (a) and some examples (b), (c), and (d) of pionic radiative corrections to this.

expressed phenomenologically in terms of the amplitudes for Λ decay and for pion-nucleon coupling. In physical terms, they represent the process of internal conversion of the pion field generated by the $\Lambda \rightarrow N+\pi$ decay interaction, due to the presence of a neighboring nucleon. The sum of these terms and the basic interaction (2.12) is

$$(fg)^{\frac{1}{2}}(1-\boldsymbol{\sigma}_\Lambda\cdot\boldsymbol{\sigma}_N)+\left\{\frac{\sqrt{2}G}{2m}\frac{4\pi(s_- - p_-\boldsymbol{\sigma}_\Lambda\cdot\mathbf{q})(-\boldsymbol{\sigma}_N\cdot\mathbf{q})}{\frac{1}{4}(m_\Lambda-m)^2-q^2-m_\pi{}^2}\right.$$

$$\left.+\frac{-G}{2m}P_{\Lambda N}{}^{\sigma}\frac{4\pi(s_0+p_0\boldsymbol{\sigma}_\Lambda\cdot\mathbf{q})(\boldsymbol{\sigma}_N\cdot\mathbf{q})}{\frac{1}{4}(m_\Lambda-m)^2-\mathbf{q}^2-m_\pi{}^2}\right\}. \quad (2.13)$$

where $\mathbf{q}$ is the momentum of the outgoing proton, G denotes the pion-nucleon coupling parameter ($G^2/hc \simeq 13.5$), $\boldsymbol{\sigma}_\Lambda$ is the spin vector of the initial Λ particle or final proton, $\boldsymbol{\sigma}_N$ is that of the initial proton or final neutron, and $P_{\Lambda N}{}^{\sigma}$ denotes the spin exchange operator $(1+\boldsymbol{\sigma}_\Lambda\cdot\boldsymbol{\sigma}_N)/2$. Since these last terms are coherent, as was realized by Cerulus,[17] the relative signs of (s_-,p_-) and (s_0,p_0) affect the rate computed for the nonmesonic capture process. The direct amplitude (2.12) interferes only with the p_- and p_0 terms of (2.13) in the total nonmesonic capture rate. The expression (2.13) now serves to illustrate two points:

(i) that the relative sign between s_- and s_0, and between p_- and p_0 does have physical consequences, although these cannot be computed completely in the present case, and

(ii) that the rate of nonmesonic capture generally depends on the relative spin orientation of the Λ particle and the nucleons present. The direct term (2.12) actually vanishes for a triplet spin configuration although its value is quite appreciable $[4(fg)^{\frac{1}{2}}]$ for a singlet configuration. This may be of importance in the light hypernuclei. For example, in ${}_\Lambda H^4$, the (Λp) spin orientation appears to be singlet, from the arguments given above on the decay modes of ${}_\Lambda H^4$, whereas in the mirror nucleus ${}_\Lambda He^4$, there are two protons and the (Λp) spin orientations are randomly distributed. The ratio of the nonmesonic capture process (2.11) in these systems therefore reflects the spin dependence of the amplitude for this process. To illustrate this effect, we give in Table I correction coefficients F_s and F_p for the Ruderman-Karplus calculation of the internal conversion coefficient R, as function of the (Λp) spin state. In terms of F_s, F_p, the internal conversion coefficient $R=$ (nonmesonic capture rate)/(π^--mesonic capture rate) is given by

$$R=(F_s s_-{}^2R_s+F_p p_-{}^2R_p)/(s_-{}^2+p_-{}^2), \quad (2.14)$$

where R_s, R_p are the coefficients computed for $l=0$ and 1 by Ruderman and Karplus. This estimate includes only the "internal conversion" graphs of Figs. 2(b) and 2(c) and neglects the direct term (2.12) as well as all other radiative corrections. It is intended only to illustrate the order of magnitude of these effects.

No example of nonmesonic decay for ${}_\Lambda H^4$ has yet been reported although many examples are known of nonmesonic decay for ${}_\Lambda He^4$ and ${}_\Lambda He^5$. This contrast may possibly be the result of experimental bias, since ${}_\Lambda H^4$ nonmesonic decay can lead only to a one-pronged star.

To sum up, we emphasize again the difficulty of making a quantitative estimate of the rate of nonmesonic Λ-hypernuclear decay in terms of the elementary four-fermion interaction, owing to the complications of the radiative corrections possible. However, there is one conclusion from the hypernuclear decay evidence which appears unlikely to be modified by the effect of further radiative corrections; with the estimates $R_s \simeq 1$, $R_p \simeq 17$ of Ruderman and Karplus for ${}_\Lambda He$ hypernuclei, the observed value of ~ 1.5 for $R({}_\Lambda He)$ obtained by Schlein[18] and by Silverstein[19] still requires that the amplitude p_- should be smaller than s_-. If p-channel emission were dominant relative to s-channel emission in free Λ decay, it would require very detailed cancellations to account for such a low value of R as that observed.

TABLE I. Correction factors to the internal conversion coefficient in hypernuclear decay.[a]

Spin state	(a) $\Delta T=\frac{1}{2}$ F_s	F_p	(b) Amplitudes (2.7) F_s	F_p
$S=0$	9/4	9/4	1/4	1/4
$S=1$	19/12	1/4	11/12	9/4
Spin average (appropriate for ${}_\Lambda He^4$ and ${}_\Lambda He^5$)	7/4	3/4	3/4	7/4

[a] These correction factors have been obtained in joint work with S. Eckstein.

[17] F. Cerulus, Nuovo cimento 5, 1685 (1957).

[18] P. E. Schlein, Phys. Rev. Letters 2, 220 (1959).

[19] E. Silverstein, Suppl. Nuovo cimento, 10, 41 (1959).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

The decay processes of the Σ particles,

$$\Sigma^+ \rightarrow \begin{cases} n+\pi^+, & (2.15a) \\ p+\pi^0, & (2.15b) \end{cases}$$

$$\Sigma^- \rightarrow n+\pi^-, \quad (2.15c)$$

could already take place through the interaction (2.10), since the Σ and Λ particles are strongly coupled, but probably occur also through additional four-fermion couplings involving the Σ particles. The most general such coupling would allow isotopic spin changes ΔT up to $\frac{5}{2}$ in strange particle decays. However, the isotopic spin changes may be limited to $\Delta T \leqslant \frac{3}{2}$ by considering the interaction

$$\left\{(\bar{\Sigma}^- n)+\frac{1}{\sqrt{2}}(\bar{\Sigma}^0 p)\right\}(\bar{p}n). \quad (2.16)$$

It is attractive to combine this with $(fg/2)^{\frac{1}{2}}$ times the term (2.10) to give the form

$$(fg)^{\frac{1}{2}}(\bar{Z}N)(\bar{p}n) = (fg)^{\frac{1}{2}}\left\{(\bar{\Sigma}^- n)+\left(\frac{\bar{\Lambda}+\bar{\Sigma}^0}{\sqrt{2}}p\right)\right\}(\bar{p}n), \quad (2.17)$$

where Y, Z denote the combinations

$$Z=\begin{pmatrix}(\Lambda+\Sigma^0)/\sqrt{2}\\ \Sigma^-\end{pmatrix}, \quad Y=\begin{pmatrix}\Sigma^+\\ (\Lambda-\Sigma^0)/\sqrt{2}\end{pmatrix} \quad (2.18)$$

of Σ and Λ states appropriate to the hypothesis of a universal pion-hyperon coupling, advanced by Gell-Mann[20] and by Schwinger.[21] The interaction (2.17) appears naturally on the hypothesis that the strangeness nonconserving current J_μ' consists of the terms

$$J_\mu' = f^{\frac{1}{2}}(\bar{Z}\gamma_\mu(1+\gamma_5)N+\bar{\Xi}\gamma_\mu(1+\gamma_5)Y), \quad (2.19)$$

each of which is a spinor under isotopic spin rotations. The hypothesis of a $T=\frac{1}{2}$ form for the strangeness nonconserving current has been put forward especially by Okubo *et al.*[22]

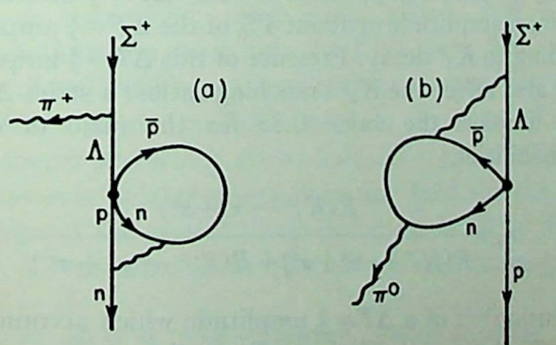


FIG. 3. Some of the lowest order graphs which lead to the Σ^+ decay processes through the four-fermion interaction $(\bar{Z}N)(\bar{N}N)$.

[20] M. Gell-Mann, Phys. Rev. **106**, 1296 (1957).
[21] J. S. Schwinger, Ann. Phys. **2**, 407 (1957).
[22] Okubo, Marshak, Sudarshan, Teutsch, and Weinberg, Phys. Rev. **112**, 665 (1958).

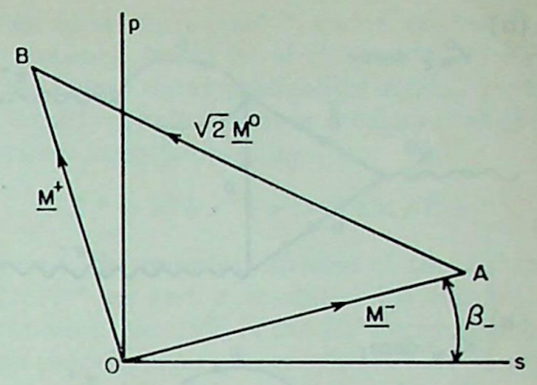


FIG. 4. The Gell-Mann—Rosenfeld diagram for representation of the matrix-elements for Σ^+ and Σ^- pionic decay modes.

The interaction current (2.19) is a $T=\frac{1}{2}$ form which corresponds to the rule $\Delta Q/\Delta s=+1$, whose empirical basis has been discussed recently by Gell-Mann and Feynman.[1] The main argument in support of this rule is the absence of evidence for $\Delta s=\pm 2$ decay interactions of strength comparable to those observed for $\Delta s=\pm 1$, since the combination of two interaction currents with $\Delta Q/\Delta s=+1$ and $\Delta Q/\Delta s=-1$ would generate a four-fermion interaction of amplitude $\sim g$ which gives rise to processes $\Delta s=\pm 2$. In accord with this rule, expression (2.19) does not include any term corresponding to decay of the Σ^+ particle. This means that the beta-decay process $\Sigma^+ \rightarrow n+e^++\bar{\nu}$ is forbidden with these interactions (however the beta decay $\Sigma^+ \rightarrow \Lambda+e^++\bar{\nu}$ is still an allowed process). The pionic processes (2.15) of Σ^+ decay are forbidden in the lowest approximation corresponding to the graphs of Fig. 1 for Λ decay, but take place through processes involving pionic radiative corrections, examples of which are shown in Fig. 3. The $\Sigma^- \rightarrow n+\pi^-$ decay, on the other hand, can take place already through the lowest order process analogous to that shown for Λ decay in Fig. 1.

For the phenomenological discussion of the decay processes (2.15), Gell-Mann and Rosenfeld[5] point out that it is convenient to represent each of the decay amplitudes for these processes as a vector $\mathbf{M}$ in an (s,p) space. With the assumption of time-reversal invariance, the smallness of $s_{\frac{1}{2}}$ and $p_{\frac{1}{2}}$ pion-nucleon phase shifts implies that the vectors $\mathbf{M}^+$, $\mathbf{M}^0$, $\mathbf{M}^-$ representing the processes (2.15) may be taken to be real. The existence of a $\Delta T=\frac{1}{2}$ rule would imply the following relationship between these real vectors,

$$\mathbf{M}^- = \mathbf{M}^+ + \sqrt{2}\mathbf{M}^0. \quad (2.20)$$

The decay rates observed for the three processes (2.15) are very nearly equal, which requires about equal magnitudes for M^-, M^+, and M^0. This requires that the vectors $\mathbf{M}^-$, $\mathbf{M}^+$, and $\sqrt{2}\mathbf{M}^0$ should form a right-angled triangle, as in Fig. 4. Carrying through the same lowest order calculation for Σ^- decay on the basis of the interaction (2.16) as was done for Λ decay leads to a value $p/s \sim 0.9$ for M^-, which would correspond to an

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

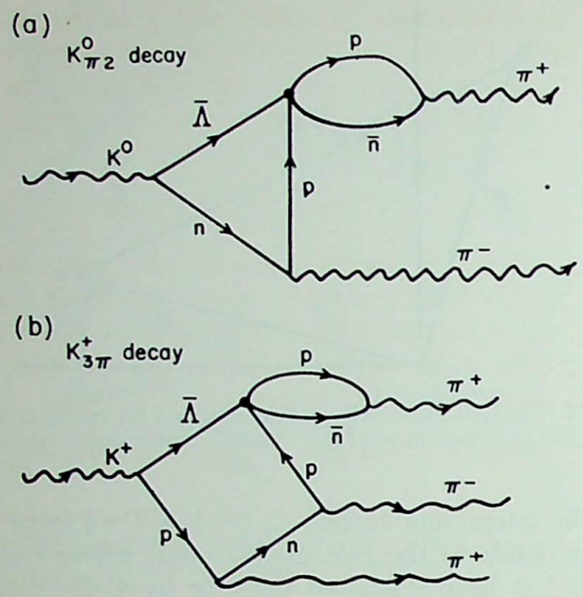


FIG. 5. Graphs illustrating the complexity of $K_{\pi 2}$ and $K_{3\pi}$ decay processes occurring through an elementary Fermi interaction.

angle $\beta_- \sim 40°$. As remarked, there is no corresponding lowest order calculation for Σ^+ decay. However the $\Delta T = \frac{1}{2}$ rule would then require that β_+ should have the value $-50°$ or $130°$, and that $\mathbf{M}^0$ should be parallel to either s or p axis.

The observations of Cool *et al.*[23] on up-down asymmetry in Σ^+ and Σ^- decay are completely opposite to these predictions, for they found that evidence for a polarization effect appears only in the (π^0+p) mode (2.15b), no polarization effect being observed for the (π^++n) mode nor for Σ^- decay. Their observations are not inconsistent with the $\Delta T = \frac{1}{2}$ rule but imply that the triangle of Fig. 4 should have its sides OA, OB quite close to the two axes, with $\mathbf{M}^0$ at an angle of order $\pm 45°$ to the axes. This failure of the lowest order calculation for Σ^- decay to agree with the data certainly casts doubt on its relevance in the case of Λ decay; there is certainly no reason to believe that the pionic radiative corrections should not have a large effect and the degree of agreement noted above for Λ decay may well be fortuitous. Alternatively, it may be that the $\Delta T = \frac{1}{2}$ rule does not hold here and that no up-down asymmetry has been observed for Σ^- decay because the relevant Σ^- production processes happen to give little polarization; if $\Delta T = \frac{3}{2}$ transitions are also effective, the present data certainly allow many other interpretations.

In terms of the interaction (2.19), $K_{\pi 2}$ and $K_{\pi 3}$ decay modes occur through even more complicated sequences of virtual processes, such as those of Fig. 5. With this situation, it is reasonable to expect the ratio of the matrix elements for $K_{\pi 2}{}^0$ and $K_{3\pi}{}^+$ decays to be of the order of magnitude $1/M$, M being the nucleon mass. Taking into account the ratio of phase space for 2π and 3π systems, the observed ratio $P(\theta_1{}^0)/P(\tau^+) = 3\times 10^3$ is in reasonable accord with this expectation. The situation for $K_{\pi 2}{}^+$ decay appears exceptional and provided the first evidence suggesting the possibility of an isotopic spin selection rule in strange particle decays, the topic discussed in the next section.

3. ISOTOPIC SPIN RELATIONSHIPS IN STRANGE PARTICLE DECAY

The possibility of a $\Delta T = \frac{1}{2}$ selection rule for strange particle decays was first suggested by the empirical data, especially by the large ratio between the partial lifetimes for the $K_{\pi 2}{}^+$ and $K_1{}^0$ decays. More recently, the possibility of isotopic spin relationships in weak decay processes has been considered in connection with the isotopic spin character of the weak interaction currents. Gell-Mann and Feynman[1] have suggested that the strangeness-conserving current involving strongly interacting particles may bear a close relationship to the $m_T = 1$ component of the isotopic spin vector for these particles; in fact, the current $(\bar{p}n)$ of expression (1.2) is already the $m_T = 1$ component of a $T = 1$ operator. On the other hand the strangeness nonconserving current (2.19) has a spinor character in isotopic spin. This would mean that the interaction (1.1) formed from $(\bar{Z}N)$ and $(\bar{p}n)$ would be a combination of $T = \frac{1}{2}$ and $\frac{3}{2}$ terms, so that $\Delta T = \frac{1}{2}$ and $\Delta T = \frac{3}{2}$ interactions would contribute to pionic decay processes for strange particles. On the other hand, the leptonic modes, which would result from combinations such as $(\bar{Z}N)(\bar{\nu}e)$, would allow only an isotopic spin change $\Delta T = \frac{1}{2}$, as far as concerns the strongly interacting particles among the decay products. The empirical evidence bearing on these possibilities is in the following.

$K_{\pi 2}$ Decay

A strict $\Delta T = \frac{1}{2}$ rule would forbid the $K_{\pi 2}{}^+$ decay since the final pions have even angular momentum and therefore isotopic spin $T = 2$ ($T = 0$ being excluded since the system has nonzero charge). Existence of $K_{\pi 2}{}^+$ decay implies the presence of some $\Delta T = \frac{3}{2}$ interaction, with an amplitude of about 4% of the $\Delta T = \frac{1}{2}$ amplitude leading to $K_1{}^0$ decay. Presence of this $\Delta T = \frac{3}{2}$ amplitude will also affect the $K_1{}^0$ branching ratios; a strict $\Delta T = \frac{1}{2}$ rule implies the value 0.33 for the ratio of decay probabilities,

$$r_0 = \frac{R(K_1{}^0 \to \pi^0+\pi^0)}{R(K_1{}^0 \to \pi^0+\pi^0)+R(K_1{}^0 \to \pi^++\pi^-)}. \tag{3.1}$$

Inclusion[24,25] of a $\Delta T = \frac{3}{2}$ amplitude which accounts for the $K_{\pi 2}{}^+$ partial lifetime allows this ratio r_0 to be between the limits

$$0.33\{1 \pm 4(2T_+/3T_0)^{\frac{1}{2}}\}, \tag{3.2}$$

[23] Cool, Cork, Cronin, and Wenzel, Bull. Am. Phys. Soc. Ser. II, 4, 83 (1959).

[24] M. Gell-Mann, Nuovo cimento 5, 758 (1957).

[25] R. H. Dalitz, Proc. Phys. Soc. (London) A69, 527 (1956).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

where T_+ and T_0 are the $K_{\pi 2}^+$ and K_1^0 partial lifetimes, that is between 0.29 and 0.37. The empirical value of this ratio is still rather uncertain. The observation of γ rays from the neutral K_1^0 mode leads to the value 0.14 ± 0.06, whereas the observation of Λ particles produced in the π^-+p reaction, but unaccompanied by K_1^0 decay in the charged mode, leads to the value† 0.30 ± 0.06.

Λ and Σ Decay

The branching ratio $R(\Lambda\to p+\pi^-)/(R(\Lambda\to p+\pi^-)+R(\Lambda\to n+\pi^0))$ is observed to have the value 0.63 ± 0.03 in close agreement with expectation $(\frac{2}{3})$ from $\Delta T=\frac{1}{2}$. As emphasized by Okubo *et al.*,[22] this agreement could also be fortuitous, since there is a certain combination of final $T=\frac{1}{2}$ and $\frac{3}{2}$ states which also corresponds to the observed ratio.

The Σ-decay situation was discussed in the foregoing. The present evidence is compatible with the $\Delta T=\frac{1}{2}$ rule but provides no strong argument for it.

$K_{3\pi}$-Decay Modes

From the spectrum observed for τ^+ decay,[26] it is reasonable to conclude that the final 3π state is predominantly symmetrical between the three pions; this is also the theoretical expectation for the decay of a K particle of spin zero.[26] There are only two totally symmetric configurations for three pions, one with $T=1$ and the other with $T=3$. If the amplitudes of these states in $K_{3\pi}$ decay are denoted by I_1 and I_3, the ratio of the τ' and τ modes in K^+ decay is then given by[25]

$$R(\tau')/R(\tau)=1.295(I_1-2I_3)^2/(2I_1+I_3)^2. \quad (3.3)$$

The observed ratio 0.32 ± 0.05 obtained from the present data[27] agrees well with the value 0.325 expected for a pure $T=1$ state; however there is also a nonzero solution for I_3/I_1 which gives the observed value. Accepting the solution $I_3/I_1\approx0$, this agreement with experiment still does not provide any support for the $\Delta T=\frac{1}{2}$ hypothesis, since both $\Delta T=\frac{1}{2}$ and $\Delta T=\frac{3}{2}$ interactions can reach only the $T=1$ state. The experimental ratio will be attained with any interaction which produces pions predominantly in a state of total symmetry and which allows only $\Delta T=\frac{1}{2}$ and $\Delta T=\frac{3}{2}$.

However this situation does not hold for the comparison of the partial lifetimes for 3π decay of the K^+ and K_2^0 mesons. With the $\Delta T=\frac{1}{2}$ rule these partial lifetimes should be equal[25,28] and $\frac{2}{5}$ of the $K_2^0\to3\pi$ decay processes should be of the mode $(\pi^++\pi^-+\pi^0)$. With the known decay probability $R(K_{3\pi}^+)=6.0\times10^6$ sec^{-1}, the $\Delta T=\frac{1}{2}$ rule therefore predicts that this mode should have decay probability

$$R(K_2^0\to\pi^++\pi^-+\pi^0)=2.4\times10^6\ \text{sec}^{-1}. \quad (3.4)$$

In contrast, the empirical lifetime of the K_2^0 meson is $8.1^{+3.2}_{-2.4}\times10^{-8}$ sec and it is stated by Bardon *et al.*[29] that not more than 15% of the K_2^0 decay events giving charged particles can represent the $(\pi^++\pi^-+\pi^0)$ mode. This corresponds to the empirical statement

$$R(K_2^0\to\pi^++\pi^-+\pi^0)<1.8(\pm0.6)\times10^6\ \text{sec}^{-1}, \quad (3.5)$$

which certainly provides a considerable overestimate for the empirical value of this decay probability since it includes events for which this interpretation is doubtful and omits the possibility of unobserved neutral modes of decay. As a result, there is some degree of disagreement at this point which may provide evidence of the influence of $\Delta T=\frac{3}{2}$ interactions on this decay mode. If I_1 and I_1' denote the contributions of $\Delta T=\frac{1}{2}$ and $\Delta T=\frac{3}{2}$ interactions, respectively, to the final $T=1$ state, the amplitude of the $T=1$ configuration is $(I_1-\frac{1}{2}I_1')$ for $K_{3\pi}^+$ decay, but (I_1+I_1') for $K_2^0\to3\pi$ decay. Generally, assuming a totally symmetric state for the final three pions, the ratio of all 3π modes for K_2^0 and K^+ particles is

$$\frac{R(K_2^0\to3\pi)}{R(K^+\to3\pi)}=\frac{(I_1+I_1')^2+\frac{3}{2}I_3^2}{(I_1-\frac{1}{2}I_1')^2+I_3^2}. \quad (3.6)$$

With $I_3=0$ (absence of $\Delta T=\frac{5}{2}$ interactions), this ratio can fall to a value of 0.5 for a $\Delta T=\frac{3}{2}$ admixture as small as $I_1'/I_1=-0.2$. The ratio $(3\pi^0)/(\pi^++\pi^-+\pi^0)$ for K_2^0 decay should still be 1.5, and it would be of interest to have this prediction of the $K_2^0\to3\pi^0$ decay rate checked, for example by exposure of a dense bubble chamber in a K_2^0 beam.

Leptonic Modes

Here the hypothesis is that $\Delta T=\frac{1}{2}$ holds for the strongly interacting products of this decay. In this sense, interactions for which $\Delta T=\frac{1}{2}$ certainly do exist, as shown by the existence of the $K_{\mu2}^+$ mode, for which the only strongly-interacting particle taking part is the K^+ meson itself; the question is whether these are the only interactions. At present the main physical consequence of this rule is a relation between the amplitudes for $(\pi\mu\nu)$ or $(\pi e\nu)$ decay for K^+ and K^0

† *Note added in proof.*—Crawford, Cresti, Douglass, Good, Kalbfleisch, Stevenson, and Ticho [Phys. Rev. Letters 2, 266 (1959)] have now obtained the value 0.32 ± 0.04. Their observation of γ rays from three K_1^0 decay events raises the former value to 0.18 ± 0.05, still rather low.

[26] See R. H. Dalitz, Repts. Progr. in Phys. 20, 163 (1957).

[27] Birge, Perkins, Peterson, Stork, and Whitehead, Nuovo cimento 4, 834 (1956); Alexander, Johnston, and O'Ceallaigh, Nuovo cimento 6, 478 (1957).

[28] This has also been pointed out by Okubo and Sudarshan (private communication). See also A. Pais and S. B. Treiman, Phys. Rev. 106, 1106 (1957).

[29] Bardon, Fuchs, Lande, Lederman, Chinowsky, and Tinlot, Phys. Rev. 110, 780 (1958); Bardon, Lande, Lederman, and Chinowsky, Ann. Phys. (N. Y.) 5, 156 (1958).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

states, namely

$$M(K^0 \to \pi^- + \mu^+ + \bar{\nu}) = \sqrt{2} M(K^+ \to \pi^0 + \mu^+ + \bar{\nu}), \quad (3.7)$$

and a similar relation for the electron modes. Okubo *et al.*[22] point out that, irrespective of whether or not time-reversal invariance holds for these interactions, this result implies a definite relation between the K_2^0 and K^+ decay rates

$$R(K_2^0 \to \pi^\pm + \mu^\mp + \nu) = 2R(K^+ \to \pi^0 + \mu^+ + \bar{\nu}), \quad (3.8)$$

and a similar relation for the electron modes. With the decay rates 3.4×10^6 sec^{-1} and 3.3×10^6 sec^{-1} for the $K_{\mu3}^+$ and K_{e3}^+ modes, as given by Gell-Mann and Rosenfeld,[5] a total decay probability of 13.5×10^6 sec^{-1} is therefore predicted for the leptonic modes of K_2^0 decay. Since there are observed some $K_2^0 \to 3\pi$-decay events (the $\Delta T = \frac{1}{2}$ rule would require a total of 6.0×10^6 sec^{-1} for charged and neutral modes) in addition to these leptonic modes, the observed K_2^0 decay probability, given as $12.3^{+5.2}_{-3.5}\times10^6$ sec^{-1} by Bardon *et al.*,[29] is somewhat lower than these remarks would suggest. However this does not necessarily vitiate the $\Delta T = \frac{1}{2}$ rule for the leptonic modes for it may simply indicate some degree of cancellation between I_1 and I_1' for the 3π modes of K_2^0 decay. The empirical lower limit for $K_2^0 \to (\pi^+ + \pi^- + \pi^0)$ events is 5% of all charged modes, and the addition of 7.5% for $K_2^0 \to 3\pi^0$ modes (required if $T = 1$ holds for the final 3π state) would then bring the predicted probability only up to 15.3×10^6 sec^{-1}, well within the experimental errors of the observed value.

To sum up, the only evidence pointing strongly to the $\Delta T = \frac{1}{2}$ selection rule consists of the long partial lifetime for the $K_{\pi2}^+$ decay and the $(\pi^- + p)/(\pi^0 + n)$ branching ratio in Λ decay. In these cases, further tests of the $\Delta T = \frac{1}{2}$ rule will be possible in the comparison of the polarization properties of the two Λ-decay modes, and in a definitive measurement of the branching ratio in K_1^0 decay. On the other hand, none of the data at present available offers any strong conflict with the requirements of the $\Delta T = \frac{1}{2}$ rule, and all of the data is in accord with the hypothesis of a $T = \frac{1}{2}$ weak interaction current associated with the strangeness nonconserving decays.

4. LEPTONIC MODES OF STRANGE PARTICLE DECAY

K-Meson Decays

The existence of a four-fermion coupling involving a hyperon-nucleon term and a leptonic term, of the same strength whether the lepton be electron or muon, and consisting only of V and A terms, is at present in qualitative accord with the evidence on leptonic modes of K decay.

First, such an interaction accounts very naturally for the absence of evidence for K_{e2}^+ decay. The foregoing assumptions lead to an interaction of the form

$$C_K[\bar{\psi}_\nu \gamma_\mu (1+\gamma_5)\psi_L]\partial\phi_K/\partial x_\mu = C_K m_L(\bar{\psi}_\nu(1+\gamma_5)\psi_L)\phi_K \quad (4.1)$$

for K_{L2} decay, which implies a $K_{e2}^+/K_{\mu2}^+$ ratio of 2.5×10^{-5}. This corresponds to the prediction $\pi_{e2}^+/\pi_{\mu2}^+ = 1.25\times10^{-4}$ for the pion, now borne out by experiment, and is to be compared with the experimental upper limit, $K_{e2}^+/K_{\mu2}^+ < 0.01$, at present.

The decay probability for the $K_{\mu2}^+$ mode (48×10^6 sec^{-1}) is only about 20% greater than that for $\pi_{\mu2}^+$ decay, despite the much greater energy release. These lifetimes correspond to a ratio $(C_K/C_\pi)^2 \sim 1/15$. However, the relation between C_K and C_π is not necessarily simple but may depend on the following factors:

(a) The parity of the K meson (relative to even parity for the Λ particle). The expressions for the intermediate baryon loops between the initial meson and the point at which the four-fermion weak interaction is effective are quite different for a scalar K meson from those for π-meson decay. Even for a pseudoscalar K meson, the graphs for π-meson and K-meson leptonic decays do not generally correspond in detail unless the Λ and Σ particles have the same parity, their virtual K-meson interactions are neglected and their pion interactions have the global symmetry, and their weak interactions are closely correlated in a similar manner.

(b) The amplitude f^1 of the strangeness nonconserving current need not be the same as that for the strangeness conserving current. The radiative corrections to the former current have quite different structure from those for the latter, so that, even if the weak interaction current strengths are the same in the approximation where the strong pion and K-meson interactions are turned off, these coupling strengths may be modified by different effective renormalizations. In fact, the experimental data on the beta decay of the hyperons (see the following) require that f should be of an order of magnitude smaller than g.

(c) The K-hyperon coupling strength G_K^2/hc appears to be about an order of magnitude weaker than the pion-nucleon coupling strength $G^2/\hbar c \simeq 13.5$.

Consider now the three-body leptonic modes $K_{\mu3}$ and K_{e3}. Assuming only V and A interactions to be effective, the matrix elements for these processes may generally be written in the form‡

$$\{Rp_{K\alpha} + S(p_{K\alpha} - p_{\pi\alpha})\}(\bar{\psi}_\nu \gamma_\alpha (1+\gamma_5)\psi_L) = \bar{\psi}_\nu\{Rm_K\gamma_0(1+\gamma_5) + Sm_L(1+\gamma_5)\}\psi_L. \quad (4.2)$$

R and S are scalar functions of $p_K \cdot p_\pi / M^2 = m_K\omega_\pi/M^2$, which are independent of m_L. If time-reversal invariance

‡ *Note added in proof.*—See, for example, A. Pais and S. B. Treiman, Phys. Rev. **105**, 1616 (1957), R. Gatto, Phys. Rev. **111**, 1426 (1958), and references cited there. A recent preprint, "Decay of Hyperons and Mesons from the Universal Fermi Interaction," by A. Fujii and M. Kawaguchi is also relevant.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

holds for both strong and weak interactions, the ratio R/S will be real.[30] In K_{e3} decay, the term proportional to m_L may be neglected and the only contribution is from the R term of (4.2). Since the variation of $p_K \cdot p_\pi / M^2$ over the allowed range in K_{L3} decay is only about 0.05 it appears a reasonable first approximation to neglect the energy dependence of R, in which case the shape predicted for the electron spectrum is unique. This spectrum[31] is shown in Fig. 6, where comparison is made with the available data[27,32] on K_{e3}^+ decay. The agreement is rather poor, but the data are subject to experimental bias whose effect is difficult to estimate. If the hypothesis of a $T=\frac{1}{2}$ weak interaction current is valid, the expression (4.2) holds for both $K_{e3}^\pm$ decay and for the K_2^0 modes $(e^\pm+\nu+\pi^\mp)$.

In $K_{\mu 3}$ decay, both R and S terms contribute and there are therefore a range of theoretical possibilities for the $K_{\mu 3}$ spectrum, of which two examples[31] are shown in Fig. 7. This figure also shows the data obtained on $K_{\mu 3}^+$ decay in the emulsion investigation of Alexander *et al.*[27] who attempted to assess the effects of various empirical biases on this distribution and who conclude that these are so many and so little understood that no detailed comparison with the theoretical distributions is justified at present. With neglect of the energy variation of R and S, the ratio of the decay probabilities

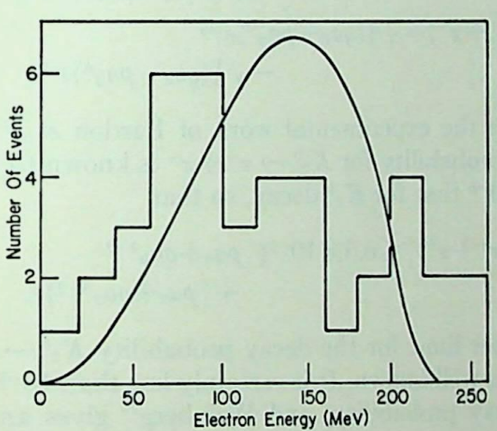


FIG. 6. The electron energy spectrum in K_{e3}^+ decay, compared with the theoretical prediction.

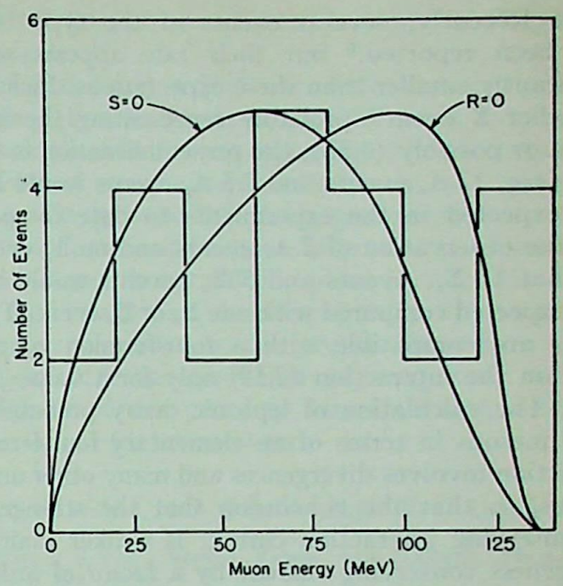


FIG. 7. The muon energy spectrum for $K_{\mu 3}^+$ decay according to Alexander *et al.*,[27] and two examples of the $K_{\mu 3}^+$ spectrum predicted by the interaction (4.2).

for K_{e3} and $K_{\mu 3}$ decay is

$$R(K_{e3})/R(K_{\mu 3}) = R^2/(0.80R^2+0.33RS+0.075S^2). \quad (4.3)$$

This ratio is limited to values less than 2.3, which does not disagree with the present empirical ratio of about unity. The ratio of the decay probabilities for K_{e3}^+ and $K_{\mu 2}^+$ modes is about 0.07, which corresponds to a value $R/C_K \sim 2/(3M)$, where M is the nucleon mass and C_K the coefficient in expression (4.1). This ratio is therefore in reasonable accord with qualitative expectation on the basis of the Fermi coupling model, since the additional pion is then emitted from the intermediate baryon pairs and the effective radius of the system is $\hbar/Mc$.

Hyperon Decays

One of the most direct consequences of Gell-Mann's tetrahedral scheme of four-fermion interactions is the prediction of a beta-decay process for the Λ and Σ^- hyperons, arising from the couplings $(\bar{\Lambda}p)(\bar{e}^+\bar{\nu})$ and $(\bar{\Sigma}^- n)(\bar{e}^+\bar{\nu})$ generated by expressions (1.1) and (1.2). With the strength $f^{\frac{1}{2}}$ of the interaction currents (2.19) equal to $g^{\frac{1}{2}}$, the expectation is that the beta decay and muon decay of the Λ particle,

$$\Lambda \rightarrow \begin{cases} p+e^-+\nu, & (4.4a) \\ p+\mu^-+\nu, & (4.4b) \end{cases}$$

should have rates 0.8% and 0.15%, respectively, of the total Λ-decay rate, and that Σ^- beta decay and muon decay

$$\Sigma^- \rightarrow \begin{cases} n+e^-+\nu, & (4.5a) \\ n+\mu^-+\nu, & (4.5b) \end{cases}$$

should have rates 5.7% and 2.5% that for normal Σ^-

[30] In the literature, it has frequently been stated that this ratio will be real as long as time-reversal invariance is valid for the strong interactions alone. When the strong interactions have this property, this conclusion is generally correct only if the K_{L3} mode arises from just one four-fermion interaction [for example, $(\bar{p}\Lambda)(\bar{\nu}L)$]. When there are several such four-fermion interactions, for example $(\bar{p}\Lambda)(\bar{\nu}L)$ and $(\bar{\Lambda}\Xi)(\bar{\nu}L)$, the K_{L3}-decay amplitude will consist of several coherent terms, thus

$$K^+ \rightarrow \left\{ \begin{matrix} \bar{\Lambda}+p \\ \bar{\Xi}^-+\Lambda \end{matrix} \right\} \rightarrow L^+ + \bar{\nu} + \pi^0,$$

and reality will hold for R/S only if the ratio of the coupling coefficients of these four-fermion interactions is also real; that is, if the weak interaction also satisfies time-reversal invariance.

[31] These K_{e3} and $K_{\mu 3}$ spectra are taken from the paper by Furuichi, Kodama, Ogawa, Sugahara, Wakasa, and Yonezawa, Progr. Theoret. Phys. 17, 89 (1957).

[32] Bruin, Holthuizen, and Jongejans, Nuovo cimento 9, 422 (1958), have collected together the data obtained with the same selection criteria.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

decay. Recently, several events of the type (4.4a) have been reported,[6] but their rate appears to be significantly smaller than these expectations. Including an earlier Σ event[33] probably representing the mode (4.5a), or possibly (4.5b), the present situation is that, with $f=g$, 12 Λ_e events and 2.5 Λ_μ events would have been expected in the experiments to date compared with the observation of 2 Λ_e events and no Λ_μ events, and that 11 Σ_e^- events and 5 Σ_μ^- events would have been expected compared with one Σ_μ or Σ_e event. These results are compatible with a four-fermion coupling based on the interaction (2.19) only for a value $f^{\frac{1}{2}}/g^{\frac{1}{2}} \sim 0.3$. The calculation of leptonic decay probabilities for K mesons in terms of an elementary four-fermion interaction involves divergences and many other uncertainties, so that the conclusion that the strangeness nonconserving interaction current is weaker than the strangeness conserving current by a factor of order 3 does not conflict with any evidence on K-meson decay.

5. TIME-REVERSAL INVARIANCE FOR WEAK INTERACTIONS

It has frequently been assumed that the weak interactions are invariant under time-reversal. In the present framework, this may be expressed as the assumption that the interaction constants for each term in the interaction current (1.2) may all be chosen to have the same phase.

There is very little direct information available at present on the validity of this assumption. In Λ decay, if the amplitudes s_- and p_- had relative phase ϕ, the expression for the polarization parameter α would have an additional factor $\cos\phi$. The experimental limitation (2.2b) shows that the angle ϕ cannot deviate by more than $\sim 45°$ from 0 or π, the values allowed with time-reversal invariance (neglecting the small pion-nucleon scattering phase shifts). No test is possible from the study of $K_{\mu 2}$-, K_{e3}-, or τ-decay modes. For $K_{\mu 3}^+$ decay, Sakurai[34] has pointed out that the violation of time-reversal for the weak interactions would generally imply that the muon polarization would generally have a component perpendicular to the (π^0,ν) plane. This possibility has yet to be examined experimentally. However, Sakurai's formulas show that, when the $K_{\mu 3}$ interaction is limited to the form (4.2), the existence of a relative phase between the coefficients R and S, which can arise if the strangeness nonconserving current (2.19) is not invariant under time-reversal, does not imply a normal component for the muon polarization.

Weinberg[35] has suggested recently that the most severe test of time-reversal at present may be the absence of 2π modes for K_2^0 decay. As pointed out by Lee *et al.*,[36] the K_1^0, K_2^0 states are generally expressible in terms of the K^0, $\bar{K}^0$ states by the relations

$$|K_1^0\rangle = p|K^0\rangle + q|\bar{K}^0\rangle, \tag{5.1a}$$

$$|K_2^0\rangle = q|K^0\rangle - p|\bar{K}^0\rangle, \tag{5.1b}$$

where p and q are generally complex numbers and $|p|^2+|q|^2=1$. If time-reversal invariance holds, which means that CP invariance is valid, then $p=q$ and both may be chosen real. Weinberg remarked that, for $j=0$, there are two final 2π states, one with $T=0$, the other $T=2$. If δ_0 and δ_2 denote the s-wave scattering phases for the pion-pion systems of $T=0$ and $T=2$, then the amplitudes for 2π decay of the K^0 system may be written

$$\langle K^0|2\pi, T=0\rangle = a_0 e^{i\delta_0}, \quad \langle K^0|2\pi, T=2\rangle = a_2 e^{i\delta_2}. \tag{5.2}$$

The corresponding amplitudes for $\bar{K}^0$ decay are directly related to these,

$$\langle \bar{K}^0|2\pi, T=0\rangle = a_0^* e^{i\delta_0}, \quad \langle \bar{K}^0|2\pi, T=2\rangle = a_2^* e^{i\delta_2}. \tag{5.3}$$

From these and Eqs. (5.1), the amplitudes for K_1^0 and K_2^0 decay to $2\pi^0$ and $\pi^+ + \pi^-$ states may then be deduced, for example,

$$\langle K_2^0|2\pi^0\rangle = \sqrt{\tfrac{1}{3}}(qa_0 - pa_0^*)e^{i\delta_0} + \sqrt{\tfrac{2}{3}}(qa_2 - pa_2^*)e^{i\delta_2}, \tag{5.4a}$$

$$\langle K_2^0|\pi^+ + \pi^-\rangle = \sqrt{\tfrac{2}{3}}(qa_0 - pa_0^*)e^{i\delta_0} - \sqrt{\tfrac{1}{3}}(qa_2 - pa_2^*)e^{i\delta_2}. \tag{5.4b}$$

From the experimental work of Bardon *et al.*,[29] the decay probability for $K_2^0 \to \pi^+ + \pi^-$ is known to be less than 10^{-5} that for K_1^0 decay, so that

$$|\langle K_2^0|\pi^+ + \pi^-\rangle| \lesssim 0.3\times 10^{-2}\{|pa_0 + qa_0^*|^2 + |pa_2 + qa_2^*|^2\}^{\frac{1}{2}}. \tag{5.5a}$$

An upper limit for the decay probability $K_2^0 \to \pi^0 + \pi^0$ is not as well known. It is certainly less than 10^{-3} of the K_1^0-decay probability, and Weinberg[35] gives an argument that this ratio is actually less than 2×10^{-4}. From this

$$|\langle K_2^0|\pi^0 + \pi^0\rangle| \lesssim 1.5\times 10^{-2}\{|pa_0 + qa_0^*|^2 + |pa_2 + qa_2^*|^2\}^{\frac{1}{2}}. \tag{5.5b}$$

If the right-hand sides of these inequalities were zero, these conditions would require, according to (5.4), that

$$a_0^*/a_0 = q/p = a_2^*/a_2, \tag{5.6}$$

and that a_0 and a_2 have the same phase, just the relationship which time-reversal invariance for the interaction $K^0 \to 2\pi$ would require.

However, owing to the uncertainty in the K_1^0 branching ratio, it is not clear exactly what restriction on the relative phases of a_2 and a_0 is implied by the empirical inequalities (5.5). If the $\Delta T = \frac{1}{2}$ rule held exactly, then the amplitude a_2 would be zero; the

[33] J. Hornbostel and E. O. Salant, Phys. Rev. **102**, 502 (1956).
[34] J. J. Sakurai, Phys. Rev. **109**, 980 (1958).
[35] S. Weinberg, Phys. Rev. **110**, 782 (1958).
[36] Lee, Oehme, and Yang, Phys. Rev. **106**, 340 (1957).

Digitized by Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

inequalities (5.5) would then imply only that $a_0^*/a_0 = q/p$, a statement about the phase of a_0 which carries no implications concerning the time-reversal invariance of the decay interaction. If the ratio of $|(pa_2+qa_2^*)|$ and $|(pa_0+qa_0^*)|$ is taken from the ratio of the $K_{\pi 2}^+$ and K_1^0 decay probabilities, then it is still true that the phases of a_0^{*2} and q/p are very close; however the limitation on the phase of a_2 is given by

$$\left|\frac{qa_2-pa_2^*}{pa_2+qa_2^*}\right| \lesssim 0.014 \left|\frac{pa_0+qa_0^*}{pa_2+qa_2^*}\right| \sim \frac{1}{3}, \qquad (5.7)$$

which involves almost no restriction on the relative phase of a_2^{*2} and q/p. This argument would lead to a significant restriction on the relative phase of a_2 and a_0 if the present upper limit for the $K_2^0 \to \pi^0+\pi^0$ decay probability were improved by an order of magnitude or if the K_1^0 branching ratio were confirmed to lie in the range 0.1–0.2, as indicated by some of the present experiments.

So far, the only serious test of the assumed property of time-reversal invariance for the weak interactions has been provided by the experiments on the beta decay of polarized neutrons reported recently by Clark *et al.*[37] and by Burgy *et al.*[38] These experiments have shown that the Fermi and Gamow-Teller matrix-elements for neutron decay do not differ in phase by more than $\pm 8°$. For the strange particle decays, the only tests which have been available to the present have been rather inconclusive or have provided only very weak evidence in support of this property for the strangeness non-conserving weak interactions.

---

[37] Clark, Robson, and Nathans, Phys. Rev. Letters **1**, 100 (1958).

[38] Burgy, Krohn, Novey, Ringo, and Telegdi, Phys. Rev. Letters **1**, 324 (1958).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

Status of Weak Interactions

MURRAY GELL-MANN

Norman Bridge Laboratory of Physics, California Institute of Technology, Pasadena, California

THE invited speakers have done an excellent job of describing the various topics within the field of weak interactions. The contributed papers have presented many fascinating results. Furthermore, what is rare these days at a physics conference, there has been plenty of discussion, which has brought out, I think, every point that was overlooked in the speeches. Evidently, there is no need for a summary.

A few words about what appear to me to be the principal questions that we would like to have answered in the future are given in the following.

Let us begin with those two beautiful and mysterious particles, the electron and the muon. The recent revolution in weak interactions has brought about measurements of the muon magnetic moment, confirming to a high degree of accuracy that the muon, like the electron, is a pure Dirac particle with electrodynamic coupling of the conventional form. It now seems clear also that the weak couplings of electron and muon are identical in form and strength, both involving the neutrino (presumably the same neutrino, although we have no way of proving that at the moment). Both electron and muon lack, so far as is known, any other interaction whatever, except the gravitational. And here appears the only known difference between them, their masses. Why do these two otherwise identical particles have different masses? No one has the slightest idea. This is perhaps the most interesting question in particle physics today. If it were not for this example, we might say that differences in mass among the elementary particles are always owing to differences in interaction. Or at least we might say that all the different particles are distinguished from one another by the values of symmetry quantum numbers. But the electron and muon seem to provide the first inkling of a radial quantum number in particle physics.

It is evidently important to refine existing measurements of the muon and electron still further, and to compare electron and muon scattering from the same target, in order to see whether the equivalence of the two particles really persists down to small distances.

The decay of the muon, involving just μ, e, and ν, offers an example of a pure weak interaction. The only known corrections are the electrodynamic ones, which are finite, have been calculated, and are easily taken into account. To within present accuracies, all our knowledge of the muon decay agrees with the very simple theory of the interaction given by the following formula for the Lagrangian density:

$$\mathfrak{L}_\mu=\sqrt{2}G\left(\bar{e}\gamma_\alpha\frac{1+\gamma_5}{\sqrt{2}}\nu\right)\left(\bar{\mu}\gamma_\alpha\frac{1+\gamma_5}{\sqrt{2}}\nu\right)^+$$

$$+\text{Hermitian conjugate (h.c.)}$$

or, for short,

$$\mathfrak{L}_\mu=\sqrt{2}G(\bar{e}\nu)(\bar{\mu}\nu)^+ + \text{h.c.} \quad (1)$$

We may describe this interaction by saying that it is the direct coupling in which e^-, μ^-, and ν appear with negative helicity only.

It seems, in fact, that in all weak interactions these particles appear with negative helicity and their antiparticles with positive helicity. It is a remarkable fact that this same grouping is followed by the law of conservation of leptons, which states that the number of leptons minus the number of antileptons is always conserved. The leptons all have negative helicity in the weak interactions, the antileptons all positive helicity.

Now there is a law of baryon conservation, too, and therefore we suspect that all baryons enter the weak interactions with the same helicity. Future experiments, suitably interpreted, will enable us to test this idea.

To return to the μ decay, it is of the highest importance to refine all experiments on the spectrum and asymmetries of μ decay in order to see whether any departures from the simple contact-interaction theory of Eq. (1) can be found. Accurate knowledge of the rate of decay is also valuable, because only in the μ decay can we be reasonably sure of measuring the true Fermi constant G, unaltered by the effects of strong couplings.

It is just these strong couplings that make β decay complicated. It is now well established that the β-decay interaction resembles that in μ decay; it is vector and axial vector and the strength is about the same. We do not, however measure the interaction Lagrangian directly, because of the strong couplings. At low momentum transfers, such as we have in nuclear β decay, we measure an *effective* Lagrangian density

$$\{\bar{n}[G_V\gamma_\alpha+(-G_A)\gamma_\alpha\gamma_5]p\}^+\left(e\gamma_\alpha\frac{1+\gamma_5}{\sqrt{2}}\nu\right)+\text{h.c.} \quad (2)$$

to be compared with

$$\{\bar{\mu}(G\gamma_\alpha+G\gamma_\alpha\gamma_5)\nu\}^+\left(\bar{e}\gamma_\alpha\frac{1+\gamma_5}{\sqrt{2}}\nu\right)+\text{h.c.}$$

as in Eq. (1) for μ decay. Here G_V and $-G_A$ are the effective or renormalized Fermi constants, containing effects of the strong interactions of nucleons.

G_V is determined experimentally from the rate of a $0\rightarrow 0$ transition such as the decay of O^{14}. Remark-

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

ably, it appears to be within one or two percent of the "pure" Fermi constant G determined from μ decay. There are two methods of measuring $-G_A/G_V$ without estimating nuclear matrix elements. One is the Argonne experiment on the asymmetry of electrons in the decay of polarized neutrons and gives $-G_A/G_V=1.25\pm0.04$. The other is a comparison of neutron and O^{14} lifetimes; it gives only the absolute value and yields

$$|-G_A/G_V|=1.19\pm0.04.$$

It is very tempting, then, to guess that the original Lagrangian for β decay is just like that for μ decay,

$$\mathfrak{L}_\beta=\sqrt{2}G(\bar{n}p)(\bar{e}\nu)^+ + \text{h.c.}, \quad (3)$$

with negative helicity for the nucleon. It is surprising, though, that the vector and axial vector renormalization factors are about 0.99 and 1.2, respectively; nearly 1 in spite of the strength of nuclear interactions. No one has any explanation for the closeness of the latter to unity. For the former, the closeness to unity is really striking, and a possible explanation has been suggested —the speculation that there is a "conserved vector current."

This speculation is suggested by an analogy between the vector weak interaction of baryons and mesons (without change of strangeness) and electromagnetism. For electromagnetism, a law of universality holds—all charged elementary particles have the same charge $\pm e$. The universality is not disturbed by strong interactions, which are present for the proton, for example, and not for the positron. The reason is that the electromagnetic current j_α is a conserved quantity: $\partial j_\alpha/\partial x_\alpha=0$.

Now the vector operator describing the nucleons in β decay is $\bar{n}\gamma_\alpha p$, which is not, in the Yukawa theory of strong interactions, a conserved quantity. It is, however, one term of a conserved quantity, the $x+iy$ component of the isotopic spin current,

$$\mathcal{J}_\alpha=\bar{n}\gamma_\alpha p+\sqrt{2}\frac{\partial\pi^+}{\partial x_\alpha}\pi^0-\sqrt{2}\pi^+\frac{\partial\pi^0}{\partial x_\alpha}$$
$$+\sqrt{2}\bar{\Sigma}^-\gamma_\alpha\Sigma^0-\sqrt{2}\bar{\Sigma}^0\gamma_\alpha\Sigma^+ +\cdots, \quad (4)$$

which obeys the conservation law $\partial\mathcal{J}_\alpha/\partial x_\alpha=0$ apart from electromagnetic effects.

If, then, in the original β-decay Lagrangian we replace $\bar{n}\gamma_\alpha(1+\gamma_5)p$ by the expression

$$\bar{n}\gamma_\alpha(1+\gamma_5)p+\sqrt{2}\frac{\partial\pi^+}{\partial x_\alpha}\pi^0-\sqrt{2}\pi^+\frac{\partial\pi^0}{\partial x_\alpha}$$
$$+\sqrt{2}\bar{\Sigma}^-\gamma_\alpha(1+\gamma_5)\Sigma^0-\sqrt{2}\bar{\Sigma}^0\gamma_\alpha(1+\gamma_5)\Sigma^+ +\cdots, \quad (5)$$

we have a theory in which G_V must equal G to within an electromagnetic correction of a percent or so. It is evidently worthwhile to refine the measurements and calculations involved in the determination of G_V/G.

It would also be good to have independent tests of the conserved current hypothesis. The simplest one is unfortunately very difficult; it involves measuring the rate of the rare decay $\pi^+\rightarrow\pi^0+e^++\nu$, never so far observed. If the conserved current hypothesis is correct, the rate* must be 0.37 ± 0.07/sec giving a branching ratio of $1.0\pm0.2\times10^{-8}$. Other β-decay theories will also lead to decay rates of this order of magnitude, but not (except accidentally) to the same number.

An easier test of the conserved current idea has been proposed, involving the ratio of the β-decay spectra of B^{12} and N^{12}. At this meeting, certain theoretical corrections have been presented that must be applied to the analysis of the results, but they do not vitiate the conclusion that the experiment can distinguish whether or not pions participate directly in the β-decay interaction as in Eq. (5). The $B^{12}-N^{12}$ experiment is being performed at the California Institute of Technology by Hilton and Sörgel, who will announce their results shortly. I would like to encourage other groups to do the same experiment, however, since it is a hard one and recent experience should have taught us that hard experiments should be repeated.

We have seen how the vector and axial vector coupling constants in β decay can be renormalized by effects of the pion cloud around the nucleon, giving, at low momentum transfer, the effective Lagrangian of Eq. (2). When we take into account finite momentum transfer to the nucleon, other pionic effects must show up, so that $G\gamma_\alpha(1+\gamma_5)$ is replaced, not simply by $G_V\gamma_\alpha+(-G_A)\gamma_\alpha\gamma_5$, but by the more complicated expression

$$G_VF_1(q^2)\gamma_\alpha+(-G_A)F_2(q^2)\gamma_\alpha\gamma_5$$
$$+AF_3(q^2)\sigma_{\alpha\beta}q_\beta+BF_4(q^2)\gamma_5q_\alpha. \quad (6)$$

Here q_α is the four-momentum transfer and the form factors $F_i(q^2)$ all have $F_i(0)=1$. Besides the renormalizations and the form factors, the pion cloud has induced two new interactions. The first, with coefficient A, we might call "weak magnetism," since it bears the same relation to the vector coupling that an anomalous Pauli moment bears to the electric charge. In fact, in the conserved current theory, the value of A can be predicted from the anomalous moments of proton and neutron. It is just this weak magnetism that is tested in the $B^{12}-N^{12}$ experiment. (The conserved current theory, by the way, also permits F_1 and F_3 to be calculated from the electromagnetic form factors of the nucleon.)

The last interaction, with coefficient B, is sometimes called the induced pseudoscalar. It can be shown that in the calculation of B by field theory, one particular type of diagram predominates, in which the nucleon radiates a virtual pion, which decays into electron and neutrino. This contribution can be evaluated exactly,

* The principal uncertainty arises from the mass difference of π^+ and π^0.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

giving a fairly reliable estimate of B, or at least its absolute value.

All these effects are barely detectible in β decay, because of the low energies involved. They may, however, be studied in the absorption of μ^- by nuclei. But how is the basic interaction of nucleons with μ and ν related to that with e and ν? The simplest assumption is that it is absolutely identical. There is very little evidence on this point from μ^- absorption itself. It has not even been proved experimentally that the interaction is V-A or that parity conservation is violated. However, the theory of identical μ and e couplings leads to the famous prediction of the branching ratio for $\pi \rightarrow e+\nu$, which has now been confirmed experimentally with an accuracy that is constantly increasing.

Pending further experimental work, we are certainly justified in supposing that the basic coupling in μ^- absorption is known. The absorption rates and other quantities may then be used to yield information about the pionic corrections of Eq. (6). For this purpose, the ideal nucleus would be the proton, but that poses tremendous experimental problems, and for a while we will have to be content with the current experiments on C^{12} and other light nuclei.

There remains the problem of calculating the absolute rate of the charged pion decay. We have heard from Professor Goldberger an excellent account of a recent attempt to do that. I think we will agree that the remarkable success of the calculation is greater than might have been expected from the approximations involved. But it is very instructive in any case, particularly the result that for large meson coupling constant g^2, the rate of pion decay is inversely proportional to g^2.

All the phenomena we have discussed so far involve the interaction of $e\nu$, $\mu\nu$, and a series of particle pairs beginning with np. The structure of the interaction is rather peculiar; pairs of particles, one neutral and one charged, interacting with one another in identical fashion. It suggests that the coupling Lagrangian is of the form $J_\alpha^+ J_\alpha$, where

$$J_\alpha = 2^{\frac{1}{2}}\sqrt{G}\left\{\bar{e}\gamma_\alpha\frac{1+\gamma_5}{\sqrt{2}}\nu+\bar{\mu}\gamma_\alpha\frac{1+\gamma_5}{\sqrt{2}}\nu+\bar{n}\gamma_\alpha\frac{1+\gamma_5}{\sqrt{2}}p+\cdots\right\}. \quad (7)$$

This idea of a current coupled to itself has certain new consequences, but they are very hard to detect by experiment. One is neutrino-electron scattering with a cross section comparable to that of neutrino absorption by nuclei. (Of course, the absorption has been detected by a coincidence technique, while the scattering gives only electron recoils.) It should be noted that the modern longitudinal or "two-component" neutrino cannot have a magnetic moment, so that if neutrino-electron scattering is ever found with the predicted cross section, it should be attributed to the direct $(e\nu)(e\nu)$ coupling.

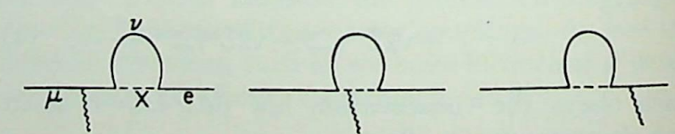
A second consequence, just as difficult to test, is the existence of a parity nonconserving nuclear force arising from the $(np)(np)$ term of the weak coupling. The violation of parity conservation should amount to something like 10^{-7} in amplitude. Present experiments are capable of measuring only one part in 10^3 or 10^4.

If we accept the form $J_\alpha^+ J_\alpha$ for the weak interactions, we are still puzzled by it. True, it is reminiscent of electromagnetism, although weaker, of short range, and not parity conserving. But the electromagnetic current j_α interacts with itself through the photon. Is there a boson to carry the weak interactions?

Let us examine the properties of such a hypothetical boson, a particle that Feynman and I like to call the *uxl* (symbol $X^\pm$). It must be a charged vector meson, say of mass M, with coupling constant (analogous to 1/137) equal to $\sqrt{2}(GM^2/4\pi)$. If M is of the order of the nucleon mass, this comes to around 10^{-6}. Processes involving creation or destruction of the *uxl* have probabilities proportional to this coupling constant, while ordinary weak processes involve the square. Thus a K particle, for example, would much rather decay into $X+\pi$ or even $X+\gamma$ than into its actual disintegration products, provided these channels involving X were open. It follows that M must be $\gtrsim m_K$.

If M is really around the nucleon mass, then the cross section for *uxl* production in, for example, nucleon-nucleon collisions of several Bev, must be around 10^{-6} of geometrical. The time of decay into, say, $e+\nu$, must be something like 10^{-18} sec. Evidently the direct observation of a particle that is rarely produced and decays immediately is a most difficult matter. If M is very much greater than 1 Bev, then the threshold for *uxl* production becomes a problem.

Fortunately, an indirect test of the existence of the *uxl* can be made. If X exists, it induces the decay $\mu \rightarrow e+\gamma$ with a rate that can be estimated. The mechanism is best described by the Feynman diagrams involved:



The rate comes out logarithmically divergent, so that we must introduce a cutoff momentum Λ. The fraction of muon decays into $e+\gamma$ then comes out

$$\Gamma(\mu \rightarrow e+\gamma)/\Gamma(\mu \rightarrow e+\nu+\bar{\nu}) = 3/8\pi\ 1/137\ f(M/\Lambda), \quad (8)$$

where for large cutoffs $\Lambda \gg M$, we have $f \rightarrow (\ln\Lambda^2/M^2)^2$. For reasonable values of the cutoff, we should expect a fraction like 10^{-3} or maybe 10^{-4}. Experimentally the branching ratio is less than 2×10^{-5}. It appears, then, that the *uxl* has flunked the only test of its existence that is available so far. (Of course, if the neutrinos

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

associated with e and μ are not identical, then this was no test at all.)

Formula (8) can also be used to describe what happens when there is no *uxl* but instead a point coupling of $(\mu\nu)$ to $(e\nu)$. Mathematically, that corresponds to the limit of infinite M with the cutoff Λ held fixed. Under these conditions $f(M/\Lambda) \to 0$ and there is no decay $\mu \to e+\gamma$.

We have not succeeded, then, in understanding the proposed form $J_\alpha{}^+ J_\alpha$ of the weak couplings. Nevertheless, I shall assume it in the remaining discussion, which concerns the weak decays of strange particles.

In order to describe decays involving a change of strangeness, we must add new terms to Eq. (7) for the current J_α, which already consists of two parts: the leptonic current $J_\alpha{}^L$ and the current $J_\alpha{}^{(1)}$ comprising $(\bar{n}p)$ and probably other terms like $(\bar{\pi}^0\pi^+)$, $(\bar{\Sigma}^-\Sigma^0)$, etc., all involving baryons or mesons, but with no change of strangeness.

The existence of decays like $K^+ \to \mu^+ + \nu$ requires us to add a current $J_\alpha{}^{(2)}$ consisting of pairs like $(\bar{\Lambda}p)$, $(\bar{\Sigma}^- n)$, etc., in which the partner with higher charge also has strangeness higher by one unit. It is easy to see that the interaction of $J_\alpha{}^{(2)}$ with $J_\alpha{}^{(1)}$ and $J_\alpha{}^L$ is sufficient to account qualitatively for all known weak decays of the strange particles.

The question now arises whether the current $J_\alpha{}^L + J_\alpha{}^{(1)} + J_\alpha{}^{(2)}$ is complete, or whether other terms of still a different character must be included. Such additional terms could be of two kinds: pairs like $(\bar{n}\Sigma^+)$ in which the partner with higher charge has strangeness lower by one unit and pairs like $(\bar{\Xi}^- n)$ in which the partners differ by two units of strangeness. Let us call the corresponding currents $J_\alpha{}^{(3)}$ and $J_\alpha{}^{(4)}$, respectively. There are a number of experimental tests of the existence of $J_\alpha{}^{(3)}$ and $J_\alpha{}^{(4)}$:

(1) The decay $\Sigma^+ \to n + e^+ + \nu$ will have appreciable probability if and only if $J^{(3)}$ is present. (Abbreviated, $\Sigma^+ \to n + e^+ + \nu \sim J^{(3)}$.)

(2) $\Xi^- \to n + e^- + \bar{\nu}$, $\Xi^0 \to p + e^- + \bar{\nu} \sim J^{(4)}$.

(3) $\Xi^- \to n + \pi^-$, $\Xi^0 \to p + \pi^- \sim J^{(3)}$ or $J^{(4)}$.

(4) If $J^{(3)}$ or $J^{(4)}$ exists, so that the weak interactions can induce strangeness changes of two units in the lowest order of G, then the transitions $K^0 \leftrightarrow \bar{K}^0$ occur with an amplitude proportional to G instead of G^2. The mass difference between $K_1{}^0$ and $K_2{}^0$ is then of order G instead of G^2, corresponding to a frequency of, say, 10^{17}/sec instead of, say, 10^{10}/sec. A beam of K^0 particles, then, will be converted immediately (that is, in something like 10^{-17} sec) into a rapidly oscillating half-and-half mixture of K^0 and $\bar{K}^0$. Thus a K^0 produced in a reaction like $\pi^- + p \to \Lambda + K^0$ will be "immediately" capable of behaving like a $\bar{K}^0$ in, for example, the reaction $\bar{K}^0 + p \to \Lambda + \pi^+$.

(5) Only if $J^{(3)}$ exists can the leptonic decay rates of $K_1{}^0$ and $K_2{}^0$ be different. If they are the same, the probability of a neutral K particle decaying into leptons during the lifetime of the $K_1{}^0$ component ($\sim 10^{-10}$ sec) is only around 10^{-3}. (It should be noted, of course, that during the first 10^{-10} sec the amplitudes for leptonic decay of $K_1{}^0$ and $K_2{}^0$ interfere with each other.)

None of these tests has yet given conclusive results. There is some evidence that a beam of K^0 particles does not immediately behave like a 50–50 mixture of K^0 and $\bar{K}^0$. If this is confirmed, we can presumably throw out both $J^{(3)}$ and $J^{(4)}$, but at the moment we cannot be sure. Let us continue, however, on the assumption that $J^{(3)}$ and $J^{(4)}$ are not present.

We must still clear up the forms of $J^{(1)}$ and $J^{(2)}$. The conserved vector current hypothesis, if we take it altogether seriously, together with the assumption that baryons always occur with the factor $1+\gamma_5$ (negative helicity), gives a definite form for $J^{(1)}$, but we are far from being able to test such an assertion today.

However, we can test a much less specific suggestion, in the form of a symmetry rule, that $J^{(1)}$ behaves like an isotopic vector and $J^{(2)}$ like an isotopic spinor. Such a rule is obeyed, for example, by the term $(\bar{n}p)$ in $J^{(1)}$ and the hypothetical term $(\bar{\Lambda}p)$ in $J^{(2)}$. If this rule is right, then apart from electromagnetic corrections, we have $|\Delta \mathbf{I}| = \frac{1}{2}$ or $\frac{3}{2}$ in nonleptonic decays and $|\Delta \mathbf{I}| = \frac{1}{2}$ in leptonic decays (with the leptons defined to carry no isotopic spin). Some experimental consequences are the following.

(1) The fraction of $K^+ \to 3\pi$ decays that yield $\pi^+ + 2\pi^0$ is $\frac{1}{5}$, assuming a totally symmetric spatial wave function.

(2) The fraction of $K_2{}^0 \to 3\pi$ decays that yield $3\pi^0$ is $\frac{3}{5}$, assuming a totally symmetric spatial wave function.

(3) The fraction of $K_1{}^0 \to 2\pi$ decays that yield $2\pi^0$ is between 0.28 and 0.38. (We make use here of the experimental rates of $K_1{}^0 \to 2\pi$ and $K^+ \to 2\pi$.)

(4) The rate of leptonic decay of $K_2{}^0$ (or of $K_1{}^0$, if we forget interference) is twice that of K^+ and the spectra, relative proportions of e and μ, etc., are all the same for $K_2{}^0$ and K^+.

All these statements seem to be roughly true except perhaps the third one, but more experiments are necessary.

We come now to some mysterious features of strange particle decay, which, if we could understand them, would probably help us to pin down the interactions further. The most striking is the fact that $K_1{}^0$ decays into 2π at a rate about 500 times faster than that of $K^+ \to 2\pi$. An explanation has been suggested in the form of an approximate rule (supposedly good to 5% or so in amplitude) that $|\Delta \mathbf{I}| = \frac{1}{2}$ in nonleptonic decays. However, it is not known how to derive any such rule in a convincing way from an interaction of the form $J_\alpha{}^+ J_\alpha$. Several ideas have been discussed, e.g., (a) to divide the strong interactions into classes, of which one is assumed to be less strong and approximately negligible. The stronger one may have a higher symmetry than charge independence, and this symmetry, together with a hypothetical symmetry of the weak

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

currents $J_\alpha^{(1)}$ and $J_\alpha^{(2)}$ may lead to something like the $|\Delta \mathbf{I}| = \frac{1}{2}$ rule. (b) to add to the interaction $J_\alpha^+ J_\alpha$ a term $K_\alpha K_\alpha$ in which the particle pairs have the same charge. The "neutral" current K_α must not contain leptons, however, since emission of pairs like $\bar{\nu}\nu$ and $\bar{\mu}e$ has never been observed. An interaction $J_\alpha^+ J_\alpha + K_\alpha K_\alpha$ can be made to yield the $|\Delta \mathbf{I}| = \frac{1}{2}$ rule.

The first suggestion seems much more attractive than the second. Actually, neither has so far resulted in a convincing theory. For a third suggestion, see below.

In any case, it is worthwhile to see whether the approximate $|\Delta \mathbf{I}| = \frac{1}{2}$ rule works experimentally. Besides the $K \to 2\pi$ situation, it explains the fact that $\Lambda \to n + \pi^0$ occurs at about half the rate of $\Lambda \to p + \pi^-$ and certain relations among the nonleptonic decays of $\Sigma^\pm$. It may be further tested by other predictions, including the following:

(1) The asymmetry of $\Lambda \to n + \pi^0$ for polarized Λ must be the same as that of $\Lambda \to p + \pi^-$.

(2) The rate of $K_2^0 \to 3\pi$ must be twice that of $K^+ \to 3\pi$. At the moment, there seems to be some difficulty with the second prediction, but again more experiments are needed.

Another mystery involves the rate of leptonic decays of hyperons. The decay $\Lambda \to p + e^- + \bar{\nu}$ has been observed, but it is reported to be very rare, occurring in much less than 1% of all Λ disintegrations. The decay $\Sigma^- \to n + e^- + \bar{\nu}$ seems also to be very rare, and has never been detected for certain. Now if the current J_2 contains terms $(\bar{\Lambda} p)$ or $(\bar{\Sigma}^- n)$, if the coefficient of each of these terms is unity, and if renormalization effects are not large (and they seem to be small in the β decay of nucleons), then the fractions of β disintegrations of Λ and Σ^- should be 1.6 and 5.6%, respectively.

While awaiting further experimental results, let us assume the discrepancy is real. If it is really large, it is unlikely to be a renormalization effect but rather an indication that the strangeness-changing weak couplings are even weaker (by a factor f, perhaps $\sim 1/20$) than those that conserve strangeness. Such a situation would permit another explanation of the approximate $|\Delta I| = \frac{1}{2}$ rule, namely that it characterizes a special set of diagrams, occurring only in nonleptonic decays, which are so much larger than other diagrams that they make up for the small factor f and give a "normal" rate for such a process as $\Lambda \to N + \pi$.

It is clear, in any case, that we have an enormous body of information about the weak decays of strange particles but that we need still more information if we are to understand them.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

Errata

Errata: Comparative Study of Potential Energy Functions for Diatomic Molecules

[Revs. Modern Phys. 29, 664 (1957)]

YATENDRA PAL VARSHNI
Department of Physics, Allahabad University, Allahabad, India

ON page 670, Eq. (40) should read

$$\alpha_e = 0.$$

In line 4, read "0" in place of "2."

On page 671, Linnett function, Eq. (55) should read

$$\alpha_e = \left[\frac{(m+1)(m-1)-l^2+3l}{3(m+1-l)}\right]\frac{6B_e^2}{\omega_e}.$$

Consequently, F values in Table II are changed as follows:

l	F
0	0.666
1	1.111
1.5	1.366
2	1.666
2.25	1.845
2.5	2.055
2.6	2.152
2.7	2.258
2.75	2.316
2.8	2.37
3.0	2.66

On page 672, Lippincott function, Eq. (61) should read

$$\alpha_e = 0.$$

In line 24, delete "As (61) shows this appears to be in error."

On page 677, Table IX, correct F values are

Linnett $(8+3l-l^2)/3(4-l)$
Lippincott 0.

On page 679, Linnett function, change the discussion for α_e as follows:

"Too high at low Δ. Suitable for CO, N_2, NO, O_2."

Errata: Experimental Results on Charge-Changing Collisions of Hydrogen and Helium Atoms and Ions at Kinetic Energies above 0.2 kev

[Revs. Modern Phys. 30, 1137 (1958)]

SAMUEL K. ALLISON
Enrico Fermi Institute for Nuclear Studies, University of Chicago, Chicago, Illinois

Page

1138 In text, above Eq. (II-1), for "$d\sigma_{if}/d\theta$ would be a function of angle" read "$d\sigma_{if}/d\Omega$ would be a function of angle."

1139 In text, above Eq. (II-5) for "and exists into high vacuum" read "and exits into high vacuum."

1140 Equation (II-12) should read

$$F_i = F_{i\infty} + [P(z,i)\exp(\pi q) + N(z,i)\exp(-\pi q)] \times \exp(-\tfrac{1}{2}\pi \textstyle\sum \sigma_{if}).$$

1141 Equation (II-28) should read

$$P(1,\bar{1}) = \frac{s+q}{2qb}[F_{0\infty}(s-q) - bF_{\bar{1}\infty}].$$

Equation (II-29) should read

$$N(1,\bar{1}) = \frac{s-q}{2qb}[F_{0\infty}(s+q) - bF_{\bar{1}\infty}].$$

Equation (II-31) should read: $dF_0/d_\pi = -(dF_1/d_\pi) = -F_{0\sigma 01} + F_{1\sigma 10}$.

1150 for "antilog 4.030–5" read "antilog (4.030–5)."

1155 Bottom line, left. Actually, Table V-6 shows that the largest observed $F_{\bar{1}\infty}$ for hydrogen is at 4 kev in argon gas.

1162 Table VI-6; for "σ_{11}" read "$\sigma_{1\bar{1}}$."

1164 Center, right, for "from Eqs. (V-1) and (V-2)" read "from equations in Sec. VA-3, page 1156."

Errata: Spin Waves

[Revs. Modern Phys. 30, 1 (1958)]

J. VAN KRANENDONK AND J. H. VAN VLECK
Harvard University, Cambridge, Massachusetts

IN Eq. (28), 0.1187 should be replaced by $\frac{1}{2}(0.1173)$.

In the third member of Eq. (32), 15/8 should be replaced by 15/32.

On the top of page 8, 0.96 $Nk/2S$ should be replaced by 0.96 $(2NkS)$.

In Eq. (97), a term $-NzJS$ should be added to the right-hand side.

In Eq. (120), 0.1187 should be replaced by $0.1173/(S_1-S_2)$ and $2JS_1S_2$ should be replaced by $4JS_1S_2$.

Below Eq. (121), $S_1S_2/S(S_1-S_2)$ should appear as twice this quantity, and $(g_1S_1-g_2S_2)\beta$ should appear multiplied by a factor $(\frac{1}{2})$.

Erratum: Dynamics of a Lattice Universe by the Schwarzschild-Cell Method

[Revs. Modern Phys. 29, 432 (1957)]

RICHARD W. LINDQUIST AND JOHN A. WHEELER
Palmer Physical Laboratory, Princeton, New Jersey

WE are indebted to Professor G. C. McVittie for pointing out to us that the printed version of Eq. (3) omits a factor.[1] It should read

$$ds^2 = [a(T)/a_0]^2\left[\frac{dr^2}{1-r^2/a_0^2} + r^2(d\theta^2+\sin^2\theta d\varphi^2)\right] - dT^2 \quad (3a)$$

$$= a^2(T)[du_1^2+du_2^2+du_3^2+du_4^2] - dT^2, \quad (3b)$$

with

$$u_1 = \sin\chi \sin\theta \sin\varphi,$$
$$u_2 = \sin\chi \sin\theta \cos\varphi,$$
$$u_3 = \sin\chi \cos\theta,$$
$$u_4 = \cos\chi,$$

and

$$r = a_0 \sin\chi. \quad (6)$$

The subsequent calculations and the conclusions drawn from them about the dynamics of a lattice universe remain unchanged.

[1] R. C. Tolman, *Relativity, Cosmology and Thermodynamics* (Clarendon Press, Oxford, England, 1934), Eqs. (149.5) and (149.10).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

Errata: Theoretical Results on Orbital Capture

[Revs. Modern Phys. 30, 1169 (1958)]

H. Brysk and M. E. Rose
Oak Ridge National Laboratory, Oak Ridge, Tennessee

In the last of Eq. (4) the first minus sign to the right of the arrow should be deleted.

In the paragraph following Eq. (4) the sentence beginning in the seventh line should read:

"The argument is weakened by the fact that the numerical value of N_0 (which appears in the cross-terms between relativistic and nonrelativistic matrix elements) has the opposite sign for capture as against positron decay; however, this appears unlikely to upset the qualitative features."

Errata: Pair Production and Bremsstrahlung in the Field of Free and Bound Electrons

[Revs. Modern Phys. 30, 354 (1958)]

J. Joseph and F. Rohrlich
Department of Physics, State University of Iowa, Iowa City, Iowa

In Eq. (2) a factor $\frac{1}{2}$ should be added to the constant factor on the right-hand side. The ratio of Eqs. (2) and (1) then becomes 2/9 as stated in the text.

In Eq. (48) $\psi_0(q)$ should be replaced by $\rho_0(q)$, the Fourier transform of $\rho_0(\mathbf{r})=\psi_0^*(\mathbf{r})\psi_0(\mathbf{r})$.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

3906
Vol. 31

REVIEWS OF
MODERN PHYSICS

VOLUME 31, NUMBER 4 OCTOBER, 1959

Internal Rotation and Microwave Spectroscopy*

CHUN C. LIN, *Department of Physics, University of Oklahoma, Norman, Oklahoma*

AND

JEROME D. SWALEN,† *Division of Pure Physics, National Research Council, Ottawa, Canada and Shell Development Company, Emeryville, California*

TABLE OF CONTENTS

1. Introduction 842
2. Hamiltonian 843
I. Potential Energy 843
II. Kinetic Energy and Hamiltonian 844
A. Internal Axis Method 844
1. Symmetric molecules 844
2. Asymmetric molecules with a plane of symmetry 846
B. Principal Axes Method 848
C. Comparison of the Two Methods 850
1. Connecting relations 849
2. Comparison by Eulerian angles. 849
(a) Principal axes method ... 849
(b) Internal axis method 850
3. High Barrier Approximations 850
I. Symmetric Molecules 851
A. Principal Axes Method 851
1. Rotational equation H_R 851
2. Torsional equation H_T 852
(a) Eigenfunctions and eigenvalues 852
(b) Harmonic oscillator approximation 853
3. Perturbation treatment 853
4. Energy levels and spectra 854
5. Reduction to a rigid rotor at a very high barrier 855
B. Internal Axis Method 855
1. Rotational equation 856
2. Torsional function 856
(a) Properties of the torsional levels 856
(b) Eigenvalues of the torsional equation 857
(c) Harmonic oscillator approximation 857
3. Energy levels and spectra 857
C. Comparison of the Two Methods 858
II. Asymmetric Molecules 858
A. Principal Axis Method 858
1. Energy levels and spectra 859
2. Calculation of energy and doublet splittings 859
3. Higher-order effects 860
(a) Third- and fourth-order perturbation terms 860
(b) Sixfold potential energy term, V_6 861
B. Internal Axis Method 861
1. Hamiltonian and matrix elements 861
2. Integrals involving the torsional functions 862
3. Diagonalization of the energy matrix 862
4. Energy levels and spectra 863
5. Splitting of the hindered rotation doublets 863
6. General case 864
7. Intermediate barrier case 864
4. Low Barrier Approximations 865
I. Free Rotation 865
A. Symmetric Molecules 865
B. Asymmetric Molecules (CH_3NO_2 type) 865
II. A Low Sixfold Barrier 866
III. A Low Threefold Barrier 867

* Supported in part by National Science Foundation, NSF-G2554.

† National Research Laboratories Postdoctorate Fellow 1956–1957.

Copyright © 1959 by the American Physical Society.
CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

5. Calculations of Barrier Heights 868
I. Analyses of Microwave Spectra 868
II. General Characteristic of the Spectra 869
III. Stark Effect 870
IV. High Barrier Case 870
V. Sample Calculation on the Spectrum of Acetaldehyde (CH_3CHO) 873
A. IAM Calculations on Acetaldehyde .. 873
B. PAM Calculations on Acetaldehyde .. 874
6. Selection Rules and Nuclear Spin Statistics 875
I. Selection Rules for Symmetric Molecules .. 875
II. Selection Rules for Asymmetric Molecules. 875
A. General Selection Rules 875
1. Molecules with a planar frame .. 875
2. Molecules with no symmetry in their framework 875
3. Molecules with a frame of C_{2v} symmetry (CH_3NO_2 type) 875
B. Selection Rules for High Barrier 876
C. Selection Rules for Low Barrier 876
1. Limiting free rotor 876
2. Low barrier 877
III. Nuclear Spin Weight 878
A. Molecules with No Symmetry 878
B. CH_3NO_2 Type Molecules 878
7. Vibration-Torsion-Rotation Interaction 878
I. Satellite Frequency Pattern 879
II. Intensity Method 881
8. Experimental Results 881
9. Conclusions 884
Appendix 1. Derivation of Kinetic Energy and Angular Momentum 885
Appendix 2. Continued Fraction 886
Appendix 3. Tabulations Related to Mathieu Equations 887
Appendix 4. Notation 887
Appendix 5. Comparison of Notation of Various Authors 889
Bibliography 889

1. INTRODUCTION

THE study of hindered internal rotation in molecules has been a subject of interest for nearly thirty years and numerous methods have been devised for investigating this phenomenon. We discuss one of these—the method of microwave spectroscopy—which has, within the last ten years, been rather extensively employed. Internal rotation interacts with the over-all rotation of a molecule and produces certain effects in its rotational spectrum which can be observed in the microwave region under high resolution. Each rotational transition exhibits a fine structure, the complexity of which depends on the height of the potential barrier hindering the internal rotation. An analysis of this fine structure leads to an accurate evaluation of the potential barrier. Considerable work has been devoted to the studies of the microwave spectra of molecules in which one group may rotate internally with respect to another. This review gives an account of the past work in this field, including certain details of the methods employed.

Potential barriers are presumably caused by the interactions of two groups of electrons and nuclei. In principle, it should be possible to determine the barrier heights from straightforward quantum-mechanical calculations. The mathematical complexity of such a treatment, however, is so great that a rigorous computation seems highly impractical at present. An alternative approach, which is perhaps somewhat empirical, is to try to describe the origin of the barriers in terms of the forces which appear in the study of intermolecular interactions, such as Van der Waals forces and resonance forces. Although many such analyses have been published, none of these results is completely satisfying. With the new data on the barrier heights in various molecules, it may be possible ultimately to formulate a simple theory of the origin of barriers which not only accounts for all the known results but also may serve to make reliable predictions, at least in a semiquantitative way. In view of the presumed similarity in nature between the origin of the potential barriers and the general problem of intermolecular forces, a satisfactory working theory of the former may throw some light on the study of the latter. The bibliography lists a number of papers which consider the origin of potential barriers and various empirical correlations of the barrier heights with the molecular structures. A review of these papers is not covered here. A brief survey of this subject is given in two recent articles by Wilson.[017,018]‡

In addition to the method of microwave spectroscopy, various other procedures have been utilized in order to evaluate potential barriers. Such thermodynamic properties as entropy and vapor heat capacity are probably the most commonly used to calculate barrier heights. Generally, these thermodynamic methods complement the microwave methods and are applicable for molecules with high barriers, greater than say 3 kcal/mole (1000 cm^{-1}). Although the thermodynamic procedure does not usually give quite as accurate barrier heights as the best results from microwave work, the former is applicable to a much larger class of molecules since the latter is limited to molecules with a dipole moment. We do not, however, review the thermodynamic results or methods here. The readers are directed to the reviews by Wilson[119] and by Pitzer.[83]

Infrared and Raman spectra[29] have been used to determine the frequencies of the internal torsional oscillations. In principle, this would be the direct method for determining the torsional frequencies and

‡ References will be found in the Bibliography at the end of the paper.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

the potential barriers. Unfortunately, these torsional oscillations are usually inactive and lie in the very far infrared region so that their detection and assignment are usually difficult, if at all possible. Nuclear magnetic resonance[1,17,85,96] also has been used recently to give a measure of the magnitude of barriers in liquids. Barrier heights in the range of 5 to 20 kcal/mole may be determined by measuring the widths of the resonance lines as functions of temperature. The time scale of nuclear magnetic resonance is such that if the barrier height is less than about 5 kcal/mole, the internal rotation appears to be free.

The phenomenon of internal rotation is similar to the inversion in ammonia,[108] a pyramidal molecule in which the nitrogen atom may be situated at either side of the plane of the hydrogen atoms. Although a small potential barrier restricts the back-and-forth motion and causes the pyramidal structure, the nitrogen atom may move from one side of the plane to the other through the quantum-mechanical tunneling effect, which splits the doubly degenerate vibrational energy levels below the barrier into pairs of levels (one symmetric and one antisymmetric state). Transitions between these states are observed in the microwave region, and the separation is very sensitive to the height of the potential barrier.

For the case of internal rotation of molecules such as ethane, CH_3-CH_3, the situation is similar. In ammonia the nitrogen atom can move back and forth between the two equivalent positions, while in CH_3-CH_3 one methyl group can rotate into one of the three positions which are equivalent with respect to the other methyl group. If each of these configurations were considered to be independent, the torsional energy levels would all be triply degenerate. However, the tunneling effect, analogous to the case of ammonia, splits each torsional level into nondegenerate (A species of the C_3 group) and degenerate (E species) sublevels. This is shown in Fig. 1 for an assumed cosine potential function. On the left are the torsional state quantum numbers. The torsional sublevel spacing increases as the torsional energy increases.

Associated with each of the torsional sublevels, one has a set of energy levels arising from the over-all rotation of the entire molecule (referred to as over-all rotational levels). Because of the interaction between the over-all and internal rotation, the spacings of the over-all rotational levels are different in the different torsional states. A given rotational transition then appears as a number of spectral lines corresponding to the transitions in different torsional states. The separations between these lines are utilized to determine the potential barrier height.

In this review we first formulate the Hamiltonians for various molecular models. The calculations of the eigenvalues and eigenfunctions based on various approximations are given and discussed. The practical aspects of the analysis of microwave spectra are presented in Sec. 5. This section is intended for those interested in evaluating potential barriers from microwave spectra. Finally, the application of group theory, vibration-torsion-rotation interaction, and experimental results are discussed in that order. The casual reader may find it convenient to omit some of the more mathematical sections in this paper. It is recommended that he first read the sections on "Symmetric Molecules" in Secs. 2 and 3 and also Sec. 4 and Sec. 8. The notation has been standardized. A glossary is given in Appendix 4 and a comparison of the notation used by other authors is given in Appendix 5.

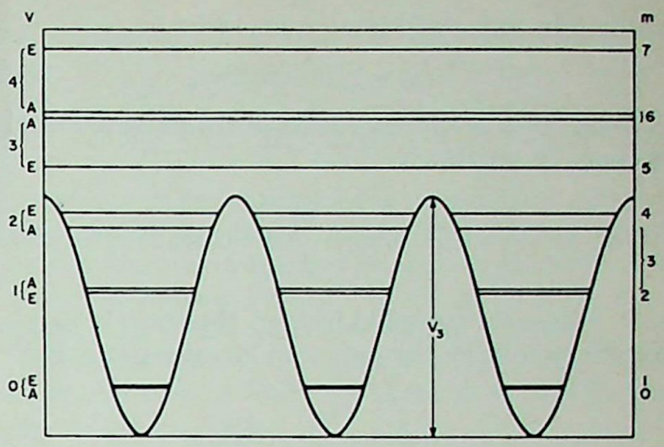


FIG. 1. Hindered rotational energy levels. A cosine potential barrier V_3 and its associated internal rotational energy levels. The torsional quantum numbers v are shown on the left ordinate. The sublevels are denoted by their symmetry under the C_3 group. The free rotational quantum numbers m are shown on the right ordinate.

2. HAMILTONIAN

In order to understand the problem of internal rotation and study its effects, the approximate Hamiltonian of the system must be known. The potential energy hindering a symmetric internal rotor is discussed first. Then the kinetic energy is written for rigid symmetric molecules and for rigid asymmetric molecules with an attached symmetric rotor. Methods of calculation of approximate eigenvalues and eigenfunctions are covered in succeeding sections.

I. Potential Energy

Since the origin of the potential barrier is not clearly understood, the only requirement that can be imposed on the potential function is that it be periodic in the relative angle α between the two parts of the molecule. Within the interval of $\alpha=0$ and $\alpha=2\pi$, the potential function must repeat itself N times, where N is the number of equivalent configurations in one complete internal revolution. For example, N is equal to three for acetaldehyde (CH_3CHO) and six for nitromethane (CH_3NO_2). In most cases the symmetry of the molecule is such that the potential function can be expressed as an even function of the angle α. The potential energy

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

can then be expanded in a cosine series as

$$V(\alpha)=\sum_k a_k \cos kN\alpha. \quad (2\text{-}1a)$$

By a shift in the reference level of the potential energy, this may be written as

$$V(\alpha)=\frac{V_3}{2}(1-\cos 3\alpha)+\frac{V_6}{2}(1-\cos 6\alpha)+\cdots \quad (2\text{-}1b)$$

for the case of a threefold barrier. However, it has been customary to take the potential function as simply[83]

$$V(\alpha)=\frac{V_3}{2}(1-\cos 3\alpha) \quad (2\text{-}2)\S$$

without rigorous justification (Fig. 1).

Serious effort has been made recently[38,39,25b] to examine the effects and magnitude of the V_6 term, and the experimental data strongly suggests that the sixfold term is indeed much smaller than the V_3 term, the ratio being the order of one-hundredth or less. If this is the case, a good approximation can then be obtained by considering only the threefold terms in the Fourier expansion of the potential energy, i.e., Eq. (2-2). The use of this simple potential function leads to solutions for the torsional wave equation in terms of Mathieu functions. In case greater accuracy is desired, the higher terms can be included and the corrections on the energy levels calculated by perturbation methods.

II. Kinetic Energy and Hamiltonian

In order to derive the kinetic energy, a model is used which consists of two rigid groups connected by a bond. *At least one of these groups is a symmetric top.* For convenience *the symmetric group is regarded as rotating internally about the bond with respect to the other group which is taken as the framework.* The entire molecule is also rotating in space. Thus there are four degrees of freedom: the three Eulerian angles θ, Φ, χ of the framework and the relative angle between the two groups, α. The method of solution is dictated by the functional form of the Hamiltonian which depends on the coordinate axes used. In the literature two methods of solution have appeared corresponding to two different frames of reference.

In that originated by Wilson[119] and Crawford[11a] *et al.*, the set of principal axes of the whole molecule is used as the coordinate system (hereafter referred to as the principal axes method or PAM). Since the rotating top possesses an axis of symmetry, the principal axes of the molecule are not altered by a change of the relative orientation of the top and the frame, and the coordinate system may be regarded as rigidly attached to the frame. The Hamiltonian function consists of three groups. The first is recognized as the energy of the rigid rotating system whose moments of inertia are *different* from those of the molecule. The second corresponds to the Hamiltonian of a simple hindered rotor with one degree of freedom. The third appears as the product of the angular momentum of the over-all rotation and the angular momentum associated with the internal motion of the top. This third group represents the interaction of the two kinds of rotation. In the solution the first two groups are taken as the unperturbed system and the cross terms, i.e., the third group, are treated as a perturbation.

In an alternative method originated by Nielsen[76] and Dennison *et al.*,[48,8] the axis about which the top executes internal rotation is chosen as one of the coordinate axes (hereafter referred to as the internal axis method or IAM). The other two axes are fixed with respect to the framework and their orientation is, in principle, arbitrary, but the choice is usually determined to some extent by the symmetry of the molecule. In this coordinate system the terms which describe the interaction between the over-all and internal rotation are considerably smaller than those from PAM. Thus the interaction terms in IAM lend themselves more readily to simple perturbation treatment. The chief disadvantage of this procedure is that, since the coordinate axes are, in general, not principal axes, the Hamiltonian is complicated by the presence of the terms containing the products of inertia.

A. Internal Axis Method

In the derivation of the Hamiltonian by the internal axis method (IAM), the case of a symmetric top (e.g., CH_3SiH_3) is treated first. An asymmetric molecule with a plane of symmetry is then discussed. Methyl alcohol (CH_3OH) and acetaldehyde (CH_3CHO) are examples of this class of molecules.

1. Symmetric molecules.[76,48]—A set of coordinate axes (a,b,c) is chosen with the c axis along the symmetry axis of the molecule through the center of mass. One of the two symmetric groups is designated the frame (a convention set up for comparison with the asymmetric case) and the other symmetric group is designated the internal rotor or top. The a and b axes can be fixed arbitrarily with respect to the frame because of the symmetric nature of this end. The kinetic energy T can then be expressed in tensor notation as (see Appendix 1)

$$T=\tfrac{1}{2}\boldsymbol{\omega}^{+}\cdot\mathbf{I}\cdot\boldsymbol{\omega}+\tfrac{1}{2}(\partial\boldsymbol{\alpha}/\partial t)^{+}\cdot\mathbf{I}_{\alpha}\cdot(\partial\boldsymbol{\alpha}/\partial t)+\tfrac{1}{2}[\boldsymbol{\omega}^{+}\cdot\mathbf{I}_{\alpha}\cdot(\partial\boldsymbol{\alpha}/\partial t)+(\partial\boldsymbol{\alpha}/\partial t)^{+}\cdot\mathbf{I}_{\alpha}\cdot\boldsymbol{\omega}], \quad (2\text{-}3)$$

or in terms of the components along the axes as

$$T=\tfrac{1}{2}I_a\omega_a^2+\tfrac{1}{2}I_b\omega_b^2+\tfrac{1}{2}I_c\omega_c^2+\tfrac{1}{2}I_\alpha\dot{\alpha}^2+I_\alpha\omega_c\dot{\alpha} \quad (2\text{-}4)$$
$$=\tfrac{1}{2}I_a\omega_a^2+\tfrac{1}{2}I_b\omega_b^2+\tfrac{1}{2}(I_c-I_\alpha)\omega_c^2+\tfrac{1}{2}I_\alpha(\omega_c+\dot{\alpha})^2,$$

where $\mathbf{I}$ is the inertial tensor and I_a, I_b, $(I_a=I_b)$, and I_c are the three moments of inertia of the whole molecule about the three coordinate axes; I_α is the moment

§ Some authors have used the form $1+\cos 3\alpha'$ which merely means a shift in α, i.e., $\alpha'=\pi-\alpha$.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

of inertia of the top about its symmetry axis, i.e., the c axis; ω_a, ω_b, and ω_c are the three components of angular velocity ω about the a, b, and c axes; and $\dot{\alpha}$ is the angular velocity of the top relative to the frame. Introduction of the momenta.

$$\begin{aligned} P_a &= \partial T/\partial\omega_a = I_a\omega_a, \\ P_b &= \partial T/\partial\omega_b = I_b\omega_b, \\ P_c &= \partial T/\partial\omega_c = I_c\omega_c + I_\alpha\dot{\alpha}, \end{aligned} \tag{2-5}$$

and

$$p = \partial T/\partial\dot{\alpha} = I_\alpha(\omega_c+\dot{\alpha}),$$

results in the classical Hamiltonian

$$H = \frac{P_a^2}{2I_a}+\frac{P_b^2}{2I_b}+\frac{P_c^2}{2(I_c-I_\alpha)} - \frac{pP_c}{(I_c-I_\alpha)}+\frac{I_c p^2}{2I_\alpha(I_c-I_\alpha)}+V(\alpha). \tag{2-6}$$

The physical meaning of P_a, P_b, and P_c is not immediately obvious. The results of Appendix 1 show that P_a, P_b, and P_c are the components of the total angular momentum (including the internal rotation) about the a, b, and c axes. Therefore they satisfy the usual commutation relations. The quantum-mechanical operators for P_a, P_b, and P_c in the coordinate representation depend only on the Eulerian angles θ, Φ, χ of the framework of the molecule but not on α (see Sec. IIC2). In this case p is the total angular momentum of the top including both the external and internal rotation. As a quantum-mechanical differential operator, p may be expressed as

$$p \rightarrow (\hbar/i)(\partial/\partial\alpha)_{\theta,\Phi,\chi}. \tag{2-7}$$

and therefore p commutes with P_a, P_b, and P_c.

At the limit of a very high potential barrier, Eq. (2-6) does not reduce readily to the case of a rigid rotor plus a simple restricted internal rotor. The reason is that the operator p defined in Eq. (2-5) contains not only the angular momentum of the internal motion of the top but also the contribution from the over-all rotation. When the internal torsion is completely frozen, the classical angular velocity $\dot{\alpha}$ becomes zero and thus p becomes $I_\alpha\omega_c = (I_\alpha/I_c)P_c$. When this expression for p is substituted into Eq. (2-6), the usual energy equation for a rigid symmetric top is recovered.

The connection between Eq. (2-6) and the Hamiltonian for the limiting case of a rigid rotor can be brought out more explicitly through the Nielsen transformation,[76]

$$\begin{aligned} p' &= p-[(I_\alpha/I_c)P_c] = [1-(I_\alpha/I_c)]I_\alpha\dot{\alpha} = rI_\alpha\dot{\alpha}, \\ P_c' &= P_c, \end{aligned} \tag{2-8}$$

which eliminates the cross product pP_c and gives

$$H = \frac{P_a^2}{2I_a}+\frac{P_b^2}{2I_b}+\frac{P_c'^2}{2I_c}+\frac{I_c p'^2}{2I_\alpha(I_c-I_\alpha)}+V(\alpha). \tag{2-9}$$

The quantity p' depends solely on $\dot{\alpha}$, so that it vanishes at the classical high barrier limit when α is a constant. Here p' is not to be taken as the angular momentum of the internal motion (which is $I_\alpha\dot{\alpha}$) but rather $rI_\alpha\dot{\alpha}$. Unlike Eq. (2-6), the coefficients of the P's in Eq. (2-9) are the actual rotational constants of the entire molecule. Furthermore p' does not commute with P_a and P_b since from Eq. (2-8),

$$p'P_a - P_a p' = i\hbar(I_\alpha/I_c)P_b. \tag{2-10}$$

On the other hand, p' does commute with $P_a^2+P_b^2$ and hence the Hamiltonian in Eq. (2-9) is separable into two parts corresponding to the over-all rotation and the internal rotation.

Equation (2-9) may also be derived through a transformation of the angular velocities. In terms of ω_c' and $\dot{\alpha}'$ defined as

$$\begin{aligned} \omega_c' &= \omega_c + (I_\alpha/I_c)\dot{\alpha}, \\ \dot{\alpha}' &= \dot{\alpha}, \end{aligned} \tag{2-11}$$

Eq. (2-4) has the form

$$T = \tfrac{1}{2}I_a(\omega_a^2+\omega_b^2)+\tfrac{1}{2}I_c\omega_c'^2+\tfrac{1}{2}rI_\alpha\dot{\alpha}'^2. \tag{2-12}$$

With the new variable ω_c' it is natural to define

$$\begin{aligned} P_c' &= (\partial T/\partial\omega_c')_{\omega_a,\omega_b,\dot{\alpha}'} = I_c\omega_c' = I_c\omega_c + I_\alpha\dot{\alpha}, \\ p' &= (\partial T/\partial\dot{\alpha}')_{\omega_a,\omega_b,\omega_c'} = [1-(I_\alpha/I_c)]I_\alpha\dot{\alpha} = rI_\alpha\dot{\alpha}. \end{aligned} \tag{2-13}$$

In addition, P_a and P_b are unchanged, i.e.,

$$\begin{aligned} P_a &= (\partial T/\partial\omega_a)_{\omega_b,\omega_c'\dot{\alpha}'} = (\partial T/\partial\omega_a)_{\omega_b,\omega_c,\dot{\alpha}}, \\ P_b &= (\partial T/\partial\omega_b)_{\omega_a,\omega_c',\dot{\alpha}'} = (\partial T/\partial\omega_b)_{\omega_a,\omega_c,\dot{\alpha}}. \end{aligned} \tag{2-14}$$

Now the momenta P_c' and p' defined in Eq. (2-13) are identical to the momenta introduced in Eq. (2-8). Substitution of Eqs. (2-13) and (2-14) into Eq. (2-12) gives Eq. (2-9).

The coordinate transformation corresponding to Eq. (2-11) and also Eq. (2-8) is simply

$$\begin{aligned} \chi' &= \chi + [(I_\alpha/I_c)\alpha] \\ &= (1/I_c)[I_\alpha\chi_2 + (I_c-I_\alpha)\chi_1], \\ \alpha' &= \alpha. \end{aligned} \tag{2-15}$$

Note that χ and α are related to the Eulerian angles of the two groups of the molecule χ_1 and χ_2 as $\chi=\chi_1$, $\alpha=\chi_2-\chi_1$. One may regard θ, Φ, χ' as the Eulerian angles for a new set of axes. This coordinate system is rotating with respect to the one attached to the framework with an angular velocity of $(I_\alpha/I_c)\dot{\alpha}$ about the axis of symmetry of the molecule. These new coordinate

‖ Here, r, the reducing factor, is an important quantity used in both the IAM and the PAM.

¶ χ is the same as Φ used in references 8 and 48.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

axes are hereafter referred to as "internal rotation axes" (called the "molecule-fixed" system by Hecht and Dennison[20]). It should be remembered that to an observer located in this frame of reference, both the framework and the top appear to be moving. The angular momentum (z component) arising from the internal motion is

$$I_\alpha(\dot{\chi}_2-\dot{\chi}')+(I_c-I_\alpha)(\dot{\chi}_1-\dot{\chi}')=0. \qquad (2\text{-}16)$$

In treating the problem of vibration-rotation interaction the moving axes are usually chosen so that the angular momentum due to the internal motion vanishes to the first approximation.[120] For the present problem it would seem proper to use the internal rotation axes as the frame of reference. By letting the three components of the total angular momentum in this coordinate system be P_a', P_b', and P_c', it follows that

$$\begin{aligned} P_a' &= (\cos\rho\alpha)P_a+(\sin\rho\alpha)P_b, \\ P_b' &= (-\sin\rho\alpha)P_a+(\cos\rho\alpha)P_b, \\ P_c' &= P_c, \end{aligned} \qquad (2\text{-}17)$$

where

$$\rho=I_\alpha/I_c. \qquad (2\text{-}18)$$

For asymmetric molecules,[20] the quantity ρ is a more complicated function of the moments of inertia:

$$\rho=\frac{(I_{bb}^2+I_{bc}^2)^{\frac{1}{2}}I_\alpha}{I_{bb}I_{cc}-I_{bc}^2}, \qquad (2\text{-}19)^{31,**}$$

where I_{bb} and I_{cc} are the moments of inertia about the b and c axis, respectively, and I_{bc} is the product of inertia [see Eq. (2-23)]. For symmetric molecules one has $I_{bc}=0$, so Eq. (2-19) reduces to Eq. (2-18).

Insertion of Eq. (2-17) into Eq. (2-9) gives

$$H=\frac{P_a'^2}{2I_a}+\frac{P_b'^2}{2I_b}+\frac{P_c'^2}{2I_c}+\frac{I_c p'^2}{2I_\alpha(I_c-I_\alpha)}+V(\alpha). \qquad (2\text{-}20)$$

Equations (2-20) and (2-9) have an identical form. This situation occurs because the two principal moments of inertia I_a and I_b of the molecule are equal. One may question the basic differences between Eqs. (2-20) and (2-9) or between P_a', P_b' and P_a, P_b. The answer is that P_a and P_b, as pointed out previously, do not commute with p' while both P_a' and P_b' do. Therefore the matrix elements (diagonal and nondiagonal) of P_a' and P_b' do not involve the quantum number associated with the operator

$$[I_c/2I_\alpha(I_c-I_\alpha)]p'^2+V(\alpha). \qquad (2\text{-}21)$$

The matrices of P_a and P_b, on the other hand, are characterized by the quantum number of the torsional operator in Eq. (2-21) as well as by the over-all rotational quantum numbers. Consequently the Heisenberg matrix elements of P_a and P_b contain both the over-all and internal rotational frequencies. The same results can be anticipated from a classical consideration. When hindered rotation is present[67] the set of the internal rotation axes of the molecule rotate in the same way as the principal axes of a rigid symmetric top, while the two groups execute rapid torsion with respect to each other. As P_a and P_b represent the components of the angular momentum along the axes fixed in one of the groups, the values of these components vary at the rate of the torsional frequency.

The cross term pP_c in Eq. (2-6) implies a coupling between the internal and over-all rotation, while according to Eq. (2-20) the internal and external motion are separable. The discrepancy arises from the different choice of the coordinate systems. When the internal rotation axes are used as the frame of reference as in Eq. (2-20), the internal torsion and over-all rotation are independent of each other and hence the Hamiltonian is separable. When the framework of the molecule is taken as the moving frame, the motion of this set of coordinate axes consists of both rotation and the rapid torsional motion. Mathematically, this is reflected by the cross term in Eq. (2-6). Because of the difference in the frames of reference in Eqs. (2-6) and (2-20), the meaning of the term "over-all rotation" is different in these two cases.

2. Asymmetric molecules with a plane of symmetry.[8,20,31,32,84]—For an asymmetric molecule with a plane of symmetry, a set of coordinate axes is chosen similar to the symmetric top case with the c axis through the center of mass of the whole molecule and parallel to the top (symmetric internal rotor) axis, and the b axis through the center of mass and lying in the plane of symmetry. The inertial tensor assumes the form

$$\begin{bmatrix} I_{aa} & 0 & 0 \\ 0 & I_{bb} & -I_{bc} \\ 0 & -I_{bc} & I_{cc} \end{bmatrix}, \qquad (2\text{-}22)$$

where I_{aa}, I_{bb}, and I_{cc} are the moments of inertia and I_{bc} is the product of inertia. The kinetic energy, according to Eq. (2-3), can then be expressed as

$$T=\tfrac{1}{2}I_{aa}\omega_a^2+\tfrac{1}{2}I_{bb}\omega_b^2+\tfrac{1}{2}I_{cc}\omega_c^2 -I_{bc}\omega_b\omega_c+\tfrac{1}{2}I_\alpha\dot{\alpha}^2+I_\alpha\omega_c\alpha. \qquad (2\text{-}23)$$

Similarly, the momenta are defined as

$$\begin{aligned} P_a &= \partial T/\partial\omega_a = I_{aa}\omega_a, \\ P_b &= \partial T/\partial\omega_b = I_{bb}\omega_b-I_{bc}\omega_c, \\ P_c &= \partial T/\partial\omega_c = I_{cc}\omega_c-I_{bc}\omega_b+I_\alpha\dot{\alpha}, \\ p &= \partial T/\partial\alpha = I_\alpha(\dot{\alpha}+\omega_c). \end{aligned} \qquad (2\text{-}24)$$

As with symmetric molecules, P_a, P_b, and P_c are again the three components of the total angular momentum.

With the substitution of Eq. (2-24) into Eq. (2-23)

** The quantity ρ is the same as Itoh's[31] λ and can be defined somewhat more simply in the principal axes coordinate system [see Eq. (2-30)].

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

the following quantum-mechanical Hamiltonian is obtained:

$$H=A_aP_a^2+B_bP_b^2+C_cP_c^2+D_{bc}(P_bP_c+P_cP_b)-2D_{bc}P_bp-2C_cP_cp+Fp^2+V(\alpha), \quad (2\text{-}25)$$

where

$$A_a=\frac{h^2}{2I_{aa}},$$

$$B_b=\frac{h^2(I_{cc}-I_\alpha)}{2(I_{bb}I_{cc}-I_\alpha I_{bb}-I_{bc}^2)}=\frac{(I_{cc}-I_\alpha)h^2}{2r(I_{bb}I_{cc}-I_{bc}^2)}=\frac{(I_{cc}-I_\alpha)h^2}{2d},$$

$$C_c=\frac{h^2I_{bb}}{2(I_{bb}I_{cc}-I_\alpha I_{bb}-I_{bc}^2)}=\frac{h^2I_{bb}}{2r(I_{bb}I_{cc}-I_{bc}^2)}=\frac{I_{bb}h^2}{2d},$$

$$D_{bc}=\frac{h^2I_{bc}}{2(I_{bb}I_{cc}-I_\alpha I_{bb}-I_{bc}^2)}=\frac{h^2I_{bc}}{2r(I_{bb}I_{cc}-I_{bc}^2)}=\frac{I_{bc}h^2}{2d},$$

$$F=\frac{h^2(I_{bb}I_{cc}-I_{bc}^2)}{2I_\alpha(I_{bb}I_{cc}-I_\alpha I_{bb}-I_{bc}^2)}=\frac{h^2}{2rI_\alpha},$$

$$r=\frac{I_{bb}I_{cc}-I_{bb}I_\alpha-I_{bc}^2}{I_{bb}I_{cc}-I_{bc}^2}=1-\frac{\lambda_y^2I_\alpha}{I_y}-\frac{\lambda_z^2I_\alpha}{I_z}, \quad \text{and}$$

$$d=I_{bb}I_{cc}-I_{bb}I_\alpha-I_{bc}^2=r(I_{bb}I_{cc}-I_{bc}^2)=rI_yI_z.$$

Here I_x, I_y, and I_z are the principal moments of inertia of the whole molecule, and λ_x, λ_y, and λ_z are the direction cosines of the internal axis with respect to the principal axes.

By following the method used for the symmetric case, we can apply the transformation in Eq. (2-8) to eliminate the coupling between over-all rotation and internal rotation. Unfortunately, this transformation does not completely remove the coupling for asymmetric molecules. The results of the transformation on H are given in Eq. (2-26),

$$H=A_aP_a^2+B_bP_b^2+C_c'P_c'^2+\frac{I_{cc}-I_\alpha}{I_{cc}}D_{bc}(P_bP_c'+P_c'P_b)-D_{bc}(P_bp'+p'P_b)-2\frac{I_{bc}}{I_{cc}}D_{bc}P_c'p'+F'p'^2+V(\alpha), \quad (2\text{-}26)$$

where

$$C_c'=\frac{\hbar^2}{2I_{cc}}\left[1+\frac{2I_{bc}^2B_b}{\hbar^2I_{cc}}\right],$$

$$F'=\frac{\hbar^2I_{cc}}{2I_\alpha(I_{cc}-I_\alpha)}+\frac{I_{bc}D_{bc}\hbar^2}{I_{cc}-I_\alpha}.$$

Equation (2-26) is essentially the same Hamiltonian employed by Burkhard and Dennison[8] to obtain the Schroedinger equation. The equivalence between Eq. (2-26) and Burkhard and Dennison's wave equation is discussed later in this section. Again p' in Eq. (2-26) does not commute with P_a and P_b.

Although the coupling between p and P_c can be eliminated for the symmetric case by the transformation (2-8), it is not removed completely for the asymmetric case but is reduced by the factor $I_{bc}^2/I_{bb}I_{cc}$. In order to minimize the coupling term the internal angular momentum must be made to vanish. This is brought about by a slightly different transformation given by Hecht and Dennison[20] or by a modification in the formulation given by Itoh.[31]

Hecht and Dennison apply three transformations to the Hamiltonian given in Eq. (2-25). First, a rotation is performed in order to eliminate the P_bp coupling. Secondly, a modified Nielsen transformation is applied to remove the P_cp coupling. These two transformations are given in Eqs. (2-27) and (2-28),

$$\begin{aligned}P_a''&=P_a,\\ P_b''&=(I_{bb}P_b-I_{bc}P_c)(I_{bb}^2+I_{bc}^2)^{-\frac{1}{2}},\\ P_c''&=(I_{bc}P_b+I_{bb}P_c)(I_{bb}^2+I_{bc}^2)^{-\frac{1}{2}},\end{aligned} \quad (2\text{-}27)$$

and

$$p'=p-\rho P_c''=p-I_\alpha(I_{bb}^2+I_{bc}^2)^{\frac{1}{2}}P_c''(I_{bb}I_{cc}-I_{bc}^2)^{-1}. \quad (2\text{-}28)$$

The transformed Hamiltonian which results from these two operations is

$$H''=A_aP_a''^2+B_b''P_b''^2+C_c''P_c''^2+D_{bc}''(P_b''P_c''+P_c''P_b'')+Fp'^2+V(\alpha), \quad (2\text{-}29)$$

where

$$B_b''=\frac{h^2}{2}I_{bb}(I_{bb}^2+I_{bc}^2)^{-1}=rI_\alpha h^2(1-r)/2\rho^2d,$$

$$C_c''=\frac{h^2}{2}\left[\frac{I_{bb}+I_{cc}}{I_{bb}I_{cc}-I_{bc}^2}-\frac{I_{bb}}{I_{bb}^2+I_{bc}^2}\right]=\frac{h^2}{2}\left(\frac{1}{I_y}+\frac{1}{I_z}\right)-B_b'',$$

$$D_{bc}''=\frac{h^2}{2}I_{bc}(I_{bb}^2+I_{bc}^2)^{-1}=(rI_\alpha)^2D_{bc}/\rho^2d=\frac{h^2}{2}\frac{\rho_y\rho_z}{\rho}\left(\frac{1}{I_y}-\frac{1}{I_z}\right),$$

$$\rho=\frac{(I_{bb}^2+I_{bc}^2)^{\frac{1}{2}}I_\alpha}{(I_{bb}I_{cc}-I_{bc}^2)}=\left[\Sigma_g\left(\frac{\lambda_gI_\alpha}{I_g}\right)^2\right]^{\frac{1}{2}}, \quad (g=x,y,z),$$

$$\rho_g=\lambda_gI_\alpha/I_g. \quad (2\text{-}30)$$

Finally, the transformation described in Eq. (2-17), with ρ given by the more complicated expression Eq. (2-30), is applied to Eq. (2-29). This transformation defines the "internal rotational axes system." Referred to these axes the internal angular momentum vanishes. The resulting Hamiltonian[20] is

$$\begin{aligned}H'=&\tfrac{1}{2}(A_a+B_b'')(P_a'^2+P_b'^2)+C_c''P_c'^2+Fp'^2+V(\alpha)\\&+\tfrac{1}{2}(A_a-B_b'')[(P_a'^2-P_b'^2)\cos2\rho\alpha\\&-(P_a'P_b'+P_b'P_a')\sin2\rho\alpha]\\&+D_{bc}''[(P_b'P_c'+P_c'P_b')\cos\rho\alpha\\&+(P_a'P_c'+P_c'P_a')\sin\rho\alpha].\end{aligned} \quad (2\text{-}31)$$

It can be easily shown that P_a', P_b', and P_c' satisfy

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

the usual commutation rules and they all commute with p'.

Itoh[31] has derived the Hamiltonian (2-29) in an interesting and useful way. He also expressed the Hamiltonian in a form convenient for comparison with the PAM formulation. The over-all angular velocity is divided into two parts ω_1 and ω_2, such that the angular momentum of the whole molecule associated with ω_1 balances the internal angular momentum. This is equivalent to choosing a set of axes in which the internal angular momentum vanishes. Thus one has

$$\mathbf{I}\cdot\omega_1+I_\alpha(\partial\alpha/\partial t)=0, \tag{2-32a}$$

or

$$\omega_1=-\mathbf{I}^{-1}\cdot\hat{k}(I_\alpha\dot{\alpha})=-\varrho\dot{\alpha}, \tag{2-32b}$$

where

$$\varrho=I_\alpha(\mathbf{I}^{-1}\cdot\hat{k}). \tag{2-32c}$$

Here $\mathbf{I}$ is the inertia tensor and $\hat{k}$ is a unit vector in the direction of $(\partial\alpha/\partial t)$. The components of the vector ϱ can be found from Eqs. (2-22) and (2-32c). Referred to the a, b, and c axes, one obtains

$$\rho_a=0,\quad \rho_b=I_\alpha I_{bc}(I_{bb}I_{cc}-I_{bc}^2)^{-1},$$

and

$$\rho_c=I_\alpha I_{bb}(I_{bb}I_{cc}-I_{bc}^2)^{-1}. \tag{2-33}$$

The three components of ϱ along the principal axes and the magnitude of the vector are given by Eq. (2-30). By expressing Eq. (2-3) in terms of ω, ω_1, and α, one obtains

$$T=\tfrac{1}{2}(\omega-\omega_1)^+\cdot\mathbf{I}\cdot(\omega-\omega_1)$$
$$-\tfrac{1}{2}\omega_1^+\cdot\mathbf{I}\cdot\omega_1+\tfrac{1}{2}(\partial\alpha/\partial t)^+\cdot\mathbf{I}_\alpha\cdot(\partial\alpha/\partial t). \tag{2-34}$$

The substitution of Eq. (2-32b) into Eq. (2-34) yields

$$T=\tfrac{1}{2}(\omega+\varrho\dot{\alpha})^+\cdot\mathbf{I}\cdot(\omega+\varrho\dot{\alpha})+\tfrac{1}{2}(\mathbf{I}_\alpha-\varrho^+\cdot\mathbf{I}\cdot\varrho)\dot{\alpha}^2$$
$$=\tfrac{1}{2}\textstyle\sum_{jk}I_{jk}(\omega_j+\rho_j\dot{\alpha})(\omega_k+\rho_k\dot{\alpha})+\tfrac{1}{2}rI_\alpha\dot{\alpha}^2.$$

The resulting Hamiltonian is

$$H=\frac{\hbar^2}{2}\sum_{jk}I_{jk}^{-1}P_jP_k+\frac{\hbar^2}{2}\left(\frac{1}{rI_\alpha}\right)(p-\Pi)^2+V(\alpha), \tag{2-35}$$

where

$$\Pi=\rho_bP_b+\rho_cP_c=\varrho\cdot\mathbf{P}.$$

This Hamiltonian is the same as Eq. (2-25). A rotation of the c axis to the ϱ direction (c' direction) brings the Hamiltonian to the form given in Eq. (2-29) which then can be transformed to Eq. (2-31). This transformation is not strictly necessary since the quantum-mechanical matrix elements can be determined from Eq. (2-29) as well as Eq. (2-31) with only a different torisonal wave function. This point is discussed in Sec. 3.

Finally, two other types of molecular symmetry have been considered in the literature. Molecules with two perpendicular planes of symmetry, e.g., nitromethane, have been considered by Wilson *et al.*[121] and by Tannenbaum, Johnson, Meyers, and Gwinn.[106,107] Here all the previous formulas apply with $I_{bc}=0$. Molecules with no planes of symmetry have been treated by Hecht and Dennison[20] and by Burkhard,[7] and the method is similar to that outlined in this section.

B. Principal Axes Method (PAM)[11a,23,101,119,121]

In the PAM the symmetric top Hamiltonian is given by Eq. (2-6) with the principal axes x, y, and z coincident with the a, b, and c axes. The transformation (2-8) is not applied. In this method (for an asymmetric molecule) a coordinate system x, y, and z, rigidly attached to the framework of the molecule, is used. Furthermore, this set of axes is chosen to be the principal axes of the entire molecule with the origin at the center of mass so that the principal moments of inertia I_x, I_y, and I_z†† of the molecule are used. The orientation of the principal axes relative to the framework is not altered by the rotation of the top because of its cylindrical symmetry. Therefore from Eq. (2-3), the kinetic energy for an asymmetric molecule has the form

$$T=\tfrac{1}{2}I_x\omega_x^2+\tfrac{1}{2}I_y\omega_y^2+\tfrac{1}{2}I_z\omega_z^2+I_\alpha\lambda_x\omega_x\dot{\alpha}$$
$$+I_\alpha\lambda_y\omega_y\dot{\alpha}+I_\alpha\lambda_z\omega_z\dot{\alpha}+\tfrac{1}{2}I_\alpha\dot{\alpha}^2, \tag{2-36}$$

where ω_x, ω_y, and ω_z are the components of the angular velocity along the principal axes, and λ_x, λ_y, and λ_z are the direction cosines of the symmetry axis of the top to the principal axes. In terms of the momenta defined as

$$P_i=\partial T/\partial\omega_i\quad(i=x,y,z),$$
$$p=\partial T/\partial\dot{\alpha},$$

the Hamiltonian can be written in the form[101]

$$H=A_xP_x^2+B_yP_y^2+C_zP_z^2+\tfrac{1}{2}\sum_{x,y,z,i\neq j}D_{ij}(P_iP_j+P_jP_i)$$
$$-2\sum_{x,y,z}Q_iP_ip+Fp^2+V(\alpha), \tag{2-37}$$

where

$$A_x=\frac{\hbar^2}{2I_x}\left[1+\frac{\lambda_x^2I_\alpha}{rI_x}\right]=\frac{\hbar^2}{2I_x}+F\rho_x^2,$$

and similarly for B_y and C_z by the permutation (x, y, z),

$$D_{ij}=\frac{\hbar^2}{2}\frac{\lambda_i\lambda_jI_\alpha}{rI_iI_j}=F\rho_i\rho_j\quad(i,j=x,y,z\text{ and }i\neq j),$$

$$Q_i=\frac{\hbar^2\lambda_i}{2rI_i}=F\rho_i\quad(i=x,y,z),$$

$$F=\hbar^2/2rI_\alpha,$$

$$\rho=(\rho_x^2+\rho_y^2+\rho_z^2)^{\frac{1}{2}}=\left[\sum_{g=x,y,z}\left(\frac{\lambda_gI_\alpha}{I_g}\right)^2\right]^{\frac{1}{2}},\quad \rho_g=\lambda_gI_\alpha/I_g,$$

$$r=1-\frac{\lambda_y^2I_\alpha}{I_y}-\frac{\lambda_z^2I_\alpha}{I_z}.$$

†† Note that the notation I_x, I_y, and I_z is used instead of I_{xx}, I_{yy}, and I_{zz} because of the absence of the product of inertia terms.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

Alternatively one may express the Hamiltonian[25a] as

$$H=(\hbar^2/2rI_\alpha)(p-\Pi)^2+(\hbar^2/2)\sum_g P_g^2/I_g+V(\alpha), \quad (2\text{-}38)$$

where

$$\Pi=\varrho\cdot\mathbf{P}=\sum_g\rho_g P_g=(\lambda_x I_\alpha/I_x)P_x +(\lambda_y I_\alpha/I_y)P_y+(\lambda_z I_\alpha/I_z)P_z.$$

The similarity between Eqs. (2-38) and (2-35) should be noted.

From Eq. (A1-7) it is clear that P_x, P_y, and P_z represent the three components of the total angular momentum defined in the usual way. The commutation relations between the various momenta can be summarized as

$$(P_i,P_j)=-i\hbar P_k \quad (i,j,k=x,y,z \text{ in cyclic order}),$$
$$(p,P_i)=0.$$

Furthermore, the coupling between over-all rotation and internal rotation is apparent in the Hamiltonian by the cross terms pP_i, etc.

For the molecules with a plane of symmetry in their framework all the above formulation applies with $\lambda_x=0$ (or $\lambda_y=0$).[68,38] With two planes of symmetry both of the direction cosines are set to zero.[121]

C. Comparison of the Two Methods

The chief difference between the two methods lies in the different choice of coordinate axes. Hence, the Hamiltonian functions can be transformed to each other simply through a coordinate transformation which is given in the first part of this section. In the second part the Schroedinger equations in terms of the Eulerian angles are given for each method. While it is not necessary to express the Hamiltonian and the wave function in the coordinate representation in order to find the energy levels, a comparison of the momentum operators in terms of the Eulerian angles does show more clearly the difference between the two formulations.

1. Connecting relations.—For the purpose of demonstrating the connection between the two procedures a molecule with a plane of symmetry is taken. The relation between the components of a vector $\mathbf{r}$ in the principal axes x, y, z and the set of axes a, b, c, defined in the internal axis method is

$$\begin{bmatrix} r_a \\ r_b \\ r_c \end{bmatrix}=\begin{bmatrix} 1 & 0 & 0 \\ 0 & \lambda_z & -\lambda_y \\ 0 & \lambda_y & \lambda_z \end{bmatrix}\begin{bmatrix} r_x \\ r_y \\ r_z \end{bmatrix}. \quad (2\text{-}39)$$

The various terms in these two systems are then related by the equations

$$I_{bb}=\lambda_z^2 I_y+\lambda_y^2 I_z,$$
$$I_{cc}=\lambda_z^2 I_z+\lambda_y^2 I_y,$$
$$I_{bc}=\lambda_y\lambda_z(I_z-I_y),$$
$$I_{aa}=I_x,$$

$$r=\frac{I_{bb}I_{cc}-I_{bb}I_a-I_{bc}^2}{I_{bb}I_{cc}-I_{bc}^2}=1-\frac{\lambda_y^2 I_\alpha}{I_y}-\frac{\lambda_z^2 I_\alpha}{I_z}, \quad (2\text{-}40)$$

$$\rho=\frac{(I_{bb}^2+I_{bc}^2)^{\frac{1}{2}}I_\alpha}{(I_{bb}I_{cc}-I_{bc}^2)}=\left(\frac{\lambda_y^2 I_\alpha^2}{I_y^2}+\frac{\lambda_z^2 I_\alpha^2}{I_z^2}\right)^{\frac{1}{2}}.$$

From these formulas one can prove the equivalence of Eq. (2-25) or Eq. (2-35) and Eq. (2-37) or Eq. (2-38).

2. Comparison by Eulerian angles.—(*a*) *Principal axes method:* To express the angular momentum operators in coordinate representation use is made of the relation

$$P_x=\frac{\partial T}{\partial\omega_x}=\frac{\partial\dot\theta}{\partial\omega_x}\frac{\partial T}{\partial\dot\theta}+\frac{\partial\dot\Phi}{\partial\omega_x}\frac{\partial T}{\partial\dot\Phi}+\frac{\partial\dot\chi}{\partial\omega_x}\frac{\partial T}{\partial\dot\chi}$$
$$=\frac{\partial\dot\theta}{\partial\omega_x}P_\theta+\frac{\partial\dot\Phi}{\partial\omega_x}P_\Phi+\frac{\partial\dot\chi}{\partial\omega_x}P_\chi. \quad (2\text{-}41)$$

Since ω_x, ω_y, and ω_z constitute the angular velocity of the framework of the molecule, they are related to the Eulerian angles describing the orientation of the framework through the equations

$$\omega_x=\dot\theta\sin\chi-\dot\Phi\sin\theta\cos\chi,$$
$$\omega_y=\dot\theta\cos\chi+\dot\Phi\sin\theta\sin\chi, \quad (2\text{-}42)$$
$$\omega_z=\dot\chi+\dot\Phi\cos\theta.$$

The quantum-mechanical operators for P_x, P_y, and P_z are readily obtained if one replaces p_θ by $(\hbar/i)(\partial/\partial\theta)_{\Phi\chi\alpha}$, etc. It follows that

$$\frac{iP_x}{h}=\sin\chi\left(\frac{\partial}{\partial\theta}\right)_{\Phi\chi\alpha}-\cos\chi\csc\theta\left(\frac{\partial}{\partial\Phi}\right)_{\theta\chi\alpha} +\cos\chi\cot\theta\left(\frac{\partial}{\partial\chi}\right)_{\theta\Phi\alpha}$$
$$\frac{iP_y}{h}=\cos\chi\left(\frac{\partial}{\partial\theta}\right)_{\Phi\chi\alpha}+\sin\chi\csc\theta\left(\frac{\partial}{\partial\Phi}\right)_{\theta\chi\alpha} -\sin\chi\cot\theta\left(\frac{\partial}{\partial\chi}\right)_{\theta\Phi\alpha}$$
$$\frac{iP_z}{h}=\left(\frac{\partial}{\partial\chi}\right)_{\theta\Phi\alpha}. \quad (2\text{-}43)$$

These expressions for the angular momentum are identical to those for a rigid rotor, and the customary commutation rules can thereby be established. Since $\omega_x, \omega_y, \omega_z$ have no dependence on $\dot\alpha$ and α, one can see that

$$p=(\partial T/\partial\dot\alpha)_{\omega_x,\omega_y,\omega_z}=(\partial T/\partial\dot\alpha)_{\theta\Phi\chi}, \quad (2\text{-}44)$$

and the quantum-mechanical for p is, accordingly,

$$p=\frac{h}{i}\left(\frac{\partial}{\partial\alpha}\right)_{\theta\Phi\chi}. \quad (2\text{-}45)$$

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

Inspection of Eqs. (2-43) and (2-45) confirms the commutation relation

$$pP_i - P_i p = 0, \quad i = x, y, z.$$

Equations (2-43) and (2-45) may now be substituted in the Hamiltonian in order to derive the desired wave equation.

(*b*) *Internal axis method:* The coordinate transformation from θ, Φ, χ, α to θ, Φ, χ', α', as described in Eq. (2-15), can be used to express P_a, P_b, P_c', and p' in terms of the new set of variables

$$\begin{aligned}(\partial/\partial\chi)_\alpha &= (\partial/\partial\chi')_{\alpha'} \\ (\partial/\partial\alpha)_\chi &= \rho(\partial/\partial\chi')_{\alpha'} + (\partial/\partial\alpha')_{\chi'}.\end{aligned} \tag{2-46}$$

Equation (2-46) is valid so long as the c axis is parallel to the bond connecting the framework and the top. However, after the transformation Eq. (2-27), this is no longer true and the Eulerian angles χ, α, χ', and α' are related by more complicated expressions [see Eq. (8) of reference (20)]. In order to simplify the mathematical details, only Eq. (2-46) is used.

It is now possible to write with the aid of Eqs. (2-15) and (2-46),‡‡ the following angular momenta,

$$\begin{aligned}P_a &= \cos\rho\alpha' P_a^{\chi'\alpha'} - \sin\rho\alpha' P_b^{\chi'\alpha'}, \\ P_b &= \sin\rho\alpha' P_a^{\chi'\alpha'} + \cos\rho\alpha' P_b^{\chi'\alpha'}, \\ P_c &= \frac{h}{i}\left(\frac{\partial}{\partial\chi'}\right)_{\theta\Phi\alpha'},\end{aligned} \tag{2-47}$$

where

$$P_a^{\chi'\alpha'} = \frac{h}{i}\left[\sin\chi'\left(\frac{\partial}{\partial\theta}\right)_{\Phi\chi'\alpha'} - \cos\chi'\csc\theta\left(\frac{\partial}{\partial\Phi}\right)_{\theta\chi'\alpha'} + \cos\chi'\cot\theta\left(\frac{\partial}{\partial\chi'}\right)_{\theta\Phi\alpha'}\right], \tag{2-48}$$

$$P_b^{\chi'\alpha'} = \frac{h}{i}\left[\cos\chi'\left(\frac{\partial}{\partial\theta}\right)_{\Phi\chi'\alpha'} + \sin\chi'\csc\theta\left(\frac{\partial}{\partial\Phi}\right)_{\theta\chi'\alpha'} - \sin\chi'\cot\theta\left(\frac{\partial}{\partial\chi'}\right)_{\theta\Phi\alpha}\right].$$

The two superscripts on P_a and P_b denote two of the four "independent variables" (the other two being always θ and Φ) one chooses in taking the partial derivatives. One might well recognize that $P_a^{\chi'\alpha'}$ and $P_b^{\chi'\alpha'}$ have almost the same form as P_x and P_y in Eq. (2-43) (which in our present notation, would be written as $P_x^{\chi\alpha}$ and $P_y^{\chi\alpha}$) aside from the fact that χ' and α' are used as variables for one case while χ and α are used for the other. It is important to specify the independent variables in the partial derivatives because Eq. (2-46) shows that $(\partial/\partial\alpha)_\chi$ is different from $(\partial/\partial\alpha)_{\chi'}$ and also from $(\partial/\partial\alpha')_{\chi'}$. Since P_a and P_b involve all four angular variables, the eigenfunctions of the operator

$$P_a^2/2I_a + P_b^2/2I_b + P_{c'}^2/2I_c$$

depend on the internal coordinate α', in addition to the three Eulerian angles, in a rather complex manner. For example, these functions are not identical to the asymmetric top wave function. It is apparent that the absence of the cross product between the two kinds of momentum operators in Eq. (2-29) does not signify a complete separation between the external and internal coordinates inasmuch as the P's are actually functions of α. When the momentum operators in Eq. (2-26) are written in their differential form given in Eqs. (2-47) and (2-48) [where ρ is I_a/I_{cc} from Eq. (2-8)], the Schroedinger equation is obtained. This equation was given by Burkhard and Dennison.§§

The rather unsatisfactory feature of having α' mixed in the P's can be removed by a transformation which is the inverse of Eq. (2-47), i.e., introducing P_a' and P_b' so that

$$\begin{pmatrix} P_a' \\ P_{b'} \end{pmatrix} = \begin{pmatrix} \cos\rho\alpha & \sin\rho\alpha \\ -\sin\rho\alpha & \cos\rho\alpha \end{pmatrix}\begin{pmatrix} P_a \\ P_b \end{pmatrix} = \begin{pmatrix} P_a^{\chi'\alpha'} \\ P_b^{\chi'\alpha'} \end{pmatrix}. \tag{2-49}$$

This transformation is similar to the one employed previously in Eq. (2-17) in connection with the internal rotation axes. Equation (2-49) shows that P_a' and P_b' do not contain α explicitly and thus commute with p'.

The Hamiltonian may then be written in the Schroedinger representation with the aid of Eqs. (2-48) and (2-49).

3. HIGH BARRIER APPROXIMATIONS

In Sec. 2 the Hamiltonians were derived for various coordinate systems and for molecules of various kinds of symmetry. These Hamiltonians can generally be divided into three parts: the over-all rotational part, the internal rotation torsional part, and a coupling between over-all rotation and internal rotation. Since the separation is not complete, except in the case of a symmetric type molecule, the Schroedinger equation cannot be solved exactly; perturbation theory is usually applied. For ease in computation we use a matrix formulation, and two approximate methods for diagonalizing the matrix are commonly used. The high barrier approximation is applicable to the cases where the separations of the torsional energy levels are large compared with the rotational energy separations. The other is the low barrier approximation where the problem of the free internal rotor is first solved and the barrier treated as a perturbation. Section 4 deals with this low barrier approximation.

‡‡ These may be obtained by a procedure similar to that employed in Eq. (2-41), i.e., from the relations

$$P_a = \frac{\partial\theta}{\partial\omega_a}P_\theta + \frac{\partial\Phi}{\partial\omega_a}P_\Phi + \frac{\partial\chi'}{\partial\omega_a}P_{\chi'} + \frac{\partial\alpha'}{\partial\omega_a}p, \text{ etc.}$$

§§ See p. 411 of reference 8.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

The Schroedinger equation for a high barrier can be solved by using either the PAM or IAM.

In the PAM the energy matrix is constructed in a representation in which the over-all rotation part and the internal torsion part of the Hamiltonian are separately diagonal. The cross terms (coupling terms) between the total angular momentum and the internal angular momentum are chosen as the perturbation. The basis functions are then the product of the rigid symmetric rotor wave functions and the functions of the internal angle α, known as the torsional functions which are related to the tabulated Mathieu functions. The perturbation terms are treated by the Van Vleck transformation[34] so that after this transformation the secular equation can be approximately factored into blocks corresponding to the different torsional states. Associated with each torsional state there are two different submatrices in the energy matrix and thus two sets of rotational levels. From the separation of the pair of spectral lines arising from a given rotational transition in the two sets of rotational levels, the barrier hindering internal rotation is determined.

In the IAM the Nielsen transformation (or the modified form given by Hecht and Dennison[20] or Itoh[31]) is applied to the Hamiltonian so that the coupling terms between the internal and external coordinates in the transformed Hamiltonian are completely removed or become much smaller than those which occur in the PAM, and hence can be treated by perturbation method more readily. The basis functions of the energy matrix are again taken as the product of the rigid symmetric rotor functions and the torsional functions which are the solution of a differential equation similar to the Mathieu equation. The torsional functions here are different from those in PAM because of the different boundary conditions as explained in Sec. I. B. Because of the interaction of the external and internal rotation, the torsional functions depend on the quantum number K and, in principle, the differential equation for the torsional functions should be solved for each value of K. Since it is necessary to have the torsional functions in order to evaluate the elements of the energy matrix, the determination of the torsional functions becomes one of the major difficulties of this method. Approximate formulas[20] have been derived for the matrix elements involving the torsional functions. At the high barrier approximation the infinite energy matrix can be reduced to a series of $(2J+1)\times(2J+1)$ blocks, corresponding to the case of limiting rigid rotor. Since the coordinate axes used in setting up the Hamiltonian in Eq. (2-25) are not the principal axes for the asymmetric rotor, a coordinate transformation to the principal axes should be made so as to simplify the solution of the secular equation. When this is done, each of the submatrices becomes very similar to that of a rigid rotor with some additional elements which represent the effect of the internal torsion on the over-all rotation. From the structure of the energy levels it was found that each of the rigid-rotor spectral lines is split into a doublet as also predicted by the PAM. The relations between the separations of the doublets to the barrier height have been given by Hecht and Dennison[20] and also by Lide and Mann.[62] These approximate equations are equivalent to those derived by PAM so that one has in many cases an equal choice as to which method to use: PAM or IAM. In the discussion to follow the symmetric molecule is covered first by the PAM and the IAM. With these principles in mind the effect of asymmetry in the molecular frame is introduced into both the PAM and the IAM.

I. Symmetric Molecules

A. Principal Axes Method (PAM)

As derived in the previous section [see Eq. (2-6)], the Hamiltonian for this class of molecules can be written as

$$H=H_R+H_T+H_{IR}, \tag{3-1}$$

where

$$H_R=A_x(P_x^2+P_y^2)+C_zP_z^2,$$

$$H_T=Fp^2+V(\alpha),$$

$$H_{IR}=-2C_zP_zp,$$

$$A_x=\frac{h^2}{2I_x}=\frac{h^2}{2I_y}, \tag{3-2}$$

$$C_z=\frac{h^2}{2(I_z-I_\alpha)}=\frac{h^2}{2I_z}+F\rho^2,$$

$$F=\frac{h^2}{2rI_\alpha}=\frac{h^2I_z}{2I_\alpha(I_z-I_\alpha)}.$$

Here, H_R and H_T comprise, respectively, the terms describing the over-all rotation and internal rotation, and H_{IR} represents the interaction term.‖‖ In the calculation of the energy levels of H, we take H_R+H_T as the unperturbed Hamiltonian and treat H_{IR} as the perturbation. Since H_{IR} does not vanish at the limit of infinite barrier, the effect of the perturbation is not *small*. Consequently, for molecules with an intermediate barrier, fourth-order perturbation procedure must be applied in order to obtain accurate results for the barrier heights. Nevertheless, the advantages of this approach will be seen.

1. Rotational equation H_R.—The eigenfunctions of H_R are the symmetric rotor functions and they are usually expressed as

$$S_{JKM}(\theta,\varphi)e^{iK\chi},$$

where J is the total angular momentum and K and M are the projection of J on the body-fixed and space-fixed axes, respectively. The energy and the matrix

‖‖ For a discussion of the "interaction" between over-all and internal rotation see the last paragraph of Sec. 2 II. A1.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

elements of P are

$$E=A_x J(J+1)+(C_z-A_x)K^2,$$
$$(JK|P^2|JK)=J(J+1)\hbar^2,$$
$$(JK|P_z|JK)=K\hbar.$$

The rotational constant, C_z, associated with P_z^2 is not that of the limiting rigid rotor and hence H_R does not represent the entire Hamiltonian for the case of infinite barrier height. The reduction of H into the rigid-rotor Hamiltonian for the limiting high barrier is discussed in Sec. I. A5.

2. Torsional equation H_T.—(a) Eigenfunctions and eigenvalues: The eigenfunction $U(\alpha)$ and the eigenvalues E of H_T are to be determined from the differential equation

$$[-Fd^2/d\alpha^2+V(\alpha)]U(\alpha)=EU(\alpha). \tag{3-3}$$

Physically we know that $U(\alpha)$ must be periodic in 2π. When $V(\alpha)$ is taken as a simple sinusoidal function such as $\frac{1}{2}V_3(1-\cos 3\alpha)$, the above equation can readily be transformed through the substitutions

$$\begin{aligned} 3\alpha+\pi&=2x, & s&=4V_3/9F,\\ U[(3\alpha+\pi)/2]&=M(x), & b&=4E/9F, \end{aligned} \tag{3-4}$$

to the Mathieu equation

$$d^2M(x)/dx^2+[(b-\tfrac{1}{2}s)-\tfrac{1}{2}s\cos 2x]M(x)=0. \tag{3-5}$$

The solutions of Eq. (3-5) may be expanded by Fourier series

$$M(x)=\sum_k c_k \cos kx+d_k \sin kx.$$

Since Eq. (3-5) is invariant under the operation $x\rightarrow -x$, $x\rightarrow x+\pi$, the eigenfunctions of Eq. (3-5) have one of the following forms[104]:

$$Se_{2r}(s,x)=\sum_k De_{2k}^{(2r)} \cos 2kx \qquad \text{(period } \pi\text{)} \tag{3-6a}$$

$$Se_{2r+1}(s,x)=\sum_k De_{2k+1}^{(2r+1)} \cos(2k+1)x \qquad \text{(period } 2\pi\text{)} \tag{3-6b}$$

$$So_{2r}(s,x)=\sum_k Do_{2k}^{(2r)} \sin 2kx \qquad \text{(period } \pi\text{)} \tag{3-6c}$$

$$So_{2r+1}(s,x)=\sum_k Do_{2k+1}^{(2r+1)} \sin(2k+1)x \qquad \text{(period } 2\pi\text{)}. \tag{3-6d}$$

Here s is a parameter associated with the differential equation (3-5). For each type of the functions, $Se(s,x)$ and $So(s,x)$, one finds a series of eigenfunctions and they are labeled by the superscript r. The values of b corresponding to these functions are denoted by be_{2r}, be_{2r+1}, bo_{2r}, bo_{2r+1}. When Eqs. (3-6) are substituted in Eq. (3-5), four different infinite secular equations are obtained from which the eigenvalues can be determined. This procedure can be carried out in a systematic fashion which is described in Appendix 2. The Fourier coefficients of the eigenfunctions are then obtained from the recursion relations associated with the infinite determinantal equations. The eigenvalues and the Fourier coefficients of the associated eigenfunctions can be found in the *Tables Relating to Mathieu Functions*.[104]

From the torsional equation one may readily see that since the C_3 group operations leave Eq. (3-3) invariant¶¶; the solution of this equation can be expanded in the form

$$\sum_k A_{3k} e^{i3k\alpha} \qquad (A) \tag{3-7a}$$

$$\sum_k A_{3k+1} e^{i(3k+1)\alpha} \qquad (E_1) \tag{3-7b}$$

$$\sum_k A_{3k-1} e^{i(3k-1)\alpha} \qquad (E_2) \tag{3-7c}$$

corresponding to the symmetry species A and E of the C_3 group. The first form of the expansion in Eq. (3-7) gives rise to nondegenerate eigenfunctions while the last two constitute the doubly degenerate pairs. The series of eigenvalues and eigenfunctions of the nondegenerate solutions can be obtained from the tabulated Mathieu functions with period π, Eqs. (3-6a) and (3-6c), after the slight modification as set down in Eq. (3-4). The degenerate solutions have been tabulated recently by Kilb.[36] The relations between the notations adopted here and those used in the various tabulations may be summarized as follows:

$E_{v\sigma}=9Fb_{v\sigma}/4$, $s=4V_3/9F$, $2x=3\alpha+\pi$

	A levels ($\sigma=0$)		E levels ($\sigma=\pm 1$)	
v	$b_{v\sigma}$	reference 104	$b_{v\sigma}$	reference 36
0	b_{00}	be_0	$b_{01}(b_{0-1})$	b_0
1	b_{10}	bo_2	$b_{11}(b_{1-1})$	b_{-1}
2	b_{20}	be_2	$b_{21}(b_{2-1})$	b_1
3	b_{30}	bo_4	$b_{31}(b_{3-1})$	b_{-2}
4	b_{40}	be_4	$b_{41}(b_{4-1})$	b_2
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.

(3-8)

Each member of the series of the degenerate and nondegenerate solutions is labeled by the index v with $v=0$ for the lowest eigenvalue. At very high barriers the spacings between the nondegenerate and degenerate energy levels of the torsional equation associated with a given v are much smaller than those between levels with different v (see Fig. 1). For this reason, the index v is called the torsional quantum number and the different energy levels associated with a given v are thought of as the torsional sublevels belonging to the same torsional state. The sublevels are distinguished by the index σ with $\sigma=0$ for the nondegenerate level (species A) and $\sigma=\pm 1$ for the levels of species E (E_1: $\sigma=+1$, E_2: $\sigma=-1$). In Fig. 2 the energy levels are plotted as a function of the barrier height from free internal rotation to a relatively high barrier. The solution of the free rotational energy levels and the meaning of the "m" quantum number are discussed in Sec. 4.

A solution of the torsional equation can then be written as

$$U_{v\sigma}(\alpha)=\sum_k A_{3k+\sigma^v} e^{i(3k+\sigma)\alpha}, \tag{3-9}$$

¶¶ Strictly speaking, Eq. (3-3) belongs to the C_{3v} group. However, in the present discussion, it is more convenient to consider only the subgroup C_3.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

and the eigenvalues associated with these functions are correspondingly denoted by $E_{v\sigma}$.

(*b*) *Harmonic oscillator approximation:* With a very high barrier the internal motion degenerates into small oscillations, and the potential energy $\frac{1}{2}V_3(1-\cos 3\alpha)$ may be expanded about the equilibrium point yielding $(9/4)V_3\alpha^2$. The torsional wave function in the vicinity of the equilibrium configuration can be approximated by the harmonic oscillator functions which are denoted as $H_v(\alpha)$. The corresponding eigenvalues are

$$E_v=3(V_3F)^{\frac{1}{2}}(v+\tfrac{1}{2}). \tag{3-10}$$

There are three functions of $H_v(\alpha)$; one situated at each of three potential minima. These functions are denoted by $H_v^{(1)}$, $H_v^{(2)}$, and $H_v^{(3)}$. When the tunneling effect is considered, the correct zeroth-order torsional functions are[20,99]

$$U_{v,0}(\alpha)=\frac{1}{\sqrt{3}}[H_v^{(1)}+H_v^{(2)}+H_v^{(3)}],$$

$$U_{v,1}(\alpha)=\frac{1}{\sqrt{3}}[H_v^{(1)}+\omega H_v^{(2)}+\omega^2 H_v^{(3)}], \tag{3-11}$$

$$U_{v,-1}(\alpha)=\frac{1}{\sqrt{3}}[H_v^{(1)}+\omega^2 H_v^{(2)}+\omega H_v^{(3)}],$$

where $\omega=\exp(2\pi i/3)$ and v, as before, denotes the torsional quantum number. The first function in Eq. (3-11) is nondegenerate (A levels) and the last two are degenerate (E levels).

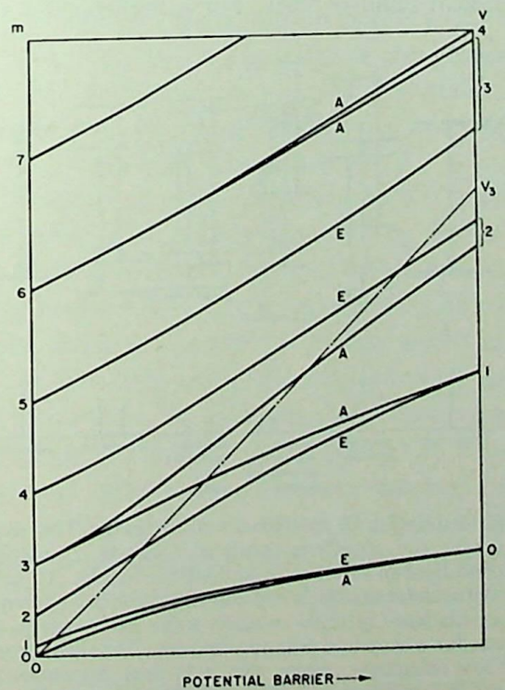


FIG. 2. The hindered internal rotational energy levels as a function of the barrier height. The free rotation quantum numbers m are at the left ordinate, i.e., a zero barrier, and the torsional quantum numbers v are on the right ordinate, i.e., a relatively high barrier. The diagonal line V_3 shows the actual height of the barrier in relation to the energy levels. Those energy levels below the top of the barrier are clearly discernible.

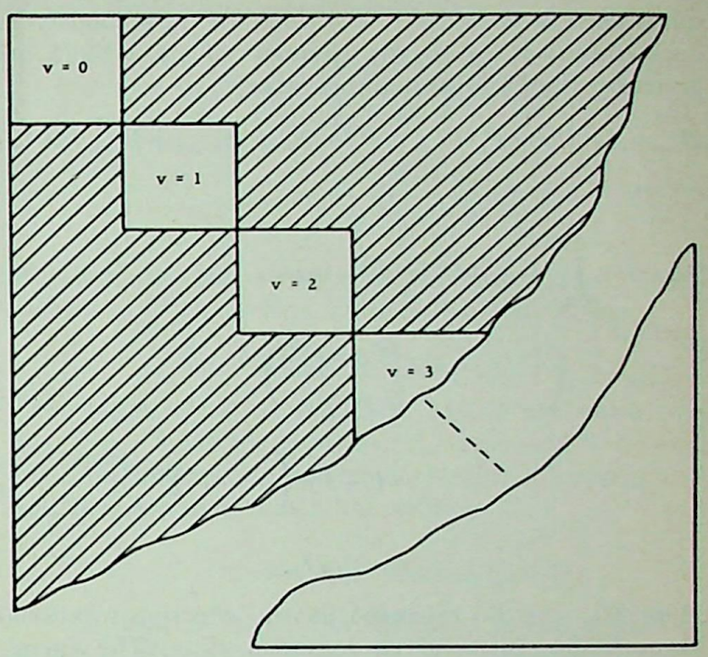


FIG. 3. The torsional matrix. Each block on the diagonal contains the over-all rotation and internal rotational matrix elements associated with that torsional level. The shaded area represents the coupling between the rotational levels of the various torsional states. With the Van Vleck transformation the matrix elements in these shaded areas are folded into the blocks on the diagonal.

3. *Perturbation treatment.*—The energy matrix for H as given in Eqs. (3-1) and (3-2) is now constructed in a representation in which H_R and H_T are diagonal. From Eq. (3-9) it is obvious that p is diagonal in the quantum number σ but not in v. Consequently, the secular equation can be grouped into blocks corresponding to different v and σ. Within each block the Hamiltonian is given by H_R+H_T, and blocks of different v are connected by the matrix elements of H_{IR}. If the energy separations between different torsional states are large compared to those between rotational levels of the same v, a VanVleck transformation[34] may be applied to the energy matrix to reduce the matrix elements nondiagonal in v to the second order; such matrix elements can then contribute only fourth-order terms to the energy and can usually be neglected. The elements diagonal in v are, of course, modified by this transformation and the effect of this transformation is to fold the elements not diagonal in v onto the v blocks.*** The matrix is illustrated in Fig. 3 where the shaded areas contain those matrix elements off diagonal in v which are transformed into the v blocks shown on the diagonal. Therefore, the infinite secular equation is effectively factored into separate blocks corresponding to various combinations of v and σ, and only one such

*** For this particular class of molecules, both H_R and H_{IR} are diagonal in K and v is the only nondiagonal index; consequently the Van Vleck transformation is identical to the ordinary perturbation treatment. However, the difference between these two methods becomes apparent for the asymmetric molecules.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

block need be considered at a time. To the second-order approximation the matrix elements of a v, σ block are given by the Hamiltonian operator[121]

$$H_{v\sigma}=A_z(P_x^2+P_y^2)+C_{v\sigma}P_z^2-2C_zP_zp_{v\sigma,v\sigma}+E_{v\sigma}, \quad (3\text{-}12)$$

where

$$p_{v\sigma,v'\sigma}=\int_0^{2\pi} U_{v\sigma}^*(\alpha)\frac{1}{i}\frac{d}{d\alpha}U_{v'\sigma}(\alpha)d\alpha,$$

$$C_{v\sigma}=C_z\left(1+4C_z\sum_v{}'\frac{|p_{v\sigma,v'\sigma}|^2}{E_{v\sigma}-E_{v'\sigma}}\right)$$

$$=C+F\rho^2\left(1+4F\sum_v{}'\frac{|p_{v\sigma,v'\sigma}|^2}{E_{v\sigma}-E_{v'\sigma}}\right),$$

$$c=\hbar^2/2I_z.$$

Here $H_{v\sigma}$ can be regarded as the effective rotational Hamiltonian for a given torsional state. The energy is then

$$E(JKMv\sigma)=A_zJ(J+1)+(C_{v\sigma}-A_z)K^2 -2C_zKp_{v\sigma,v\sigma}+E_{v\sigma}. \quad (3\text{-}13a)$$

The perturbation treatment can also be carried to higher order. The energy obtained from the nth order perturbation calculation can be expressed simply as[25a]

$$E_{v\sigma}=AJ(J+1)+(C-A)K^2 +F\sum_n w_{v\sigma}^{(n)}(\rho K)^n, \quad (3\text{-}13b)$$

where

$$\rho=I_\alpha/I_z,$$

$$w_{v\sigma}^{(0)}=9b_{v\sigma}/4,$$

$$w_{v\sigma}^{(1)}=-2p_{v\sigma,v\sigma},$$

$$w_{v\sigma}^{(2)}=1+(16/9)\sum_v{}'\frac{|p_{v\sigma,v'\sigma}|^2}{b_{v\sigma}-b_{v'\sigma}}, \text{ etc.}$$

The first two terms of Eq. (3-13b) are the energy of a rigid symmetric rotor while the last term, which is a power series in ρK, represents the effect of the internal torsion on the rotational energy levels.

The convergence of the power series depends mainly on the magnitude of ρK. Although the coefficients $w_{v\sigma}^{(n)}$ are strongly dependent on the barrier height (more precisely s) of the approximate form[99] $As^B\exp(-C\sqrt{s})$, they do not converge very rapidly with increasing n (see Fig. 8). Consequently the magnitude of the factor $(\rho K)^n$ must be examined and the perturbation calculation should then be carried out to the desired accuracy.

4. Energy levels and spectra.—For a given torsional state v there are two torsional sublevels characterized by the symmetry properties A and E (or $\sigma=0$ and ±1). Associated with each such torsional sublevel is a set of rotational energy levels, the structure of which is governed by the effective rotational Hamiltonian $H_{v\sigma}$.

For the nondegenerate A level the torsional functions are real and hence the diagonal elements of odd order in p vanish, i.e., $w_{v\sigma}^{(1)}=0$. The rotational Hamiltonian is given to the second order by the first two terms of Eq. (3-12). This is formally identical to the Hamiltonian of a rigid symmetric rotor with a modified rotational constant $C_{v\sigma}$. Consequently, the rotational levels associated with a nondegenerate A type torsional level (hereafter referred to as A rotational level or simply A levels) have the same structure as the levels of a rigid rotor (referred to also as the pseudo-rigid rotor levels). For the case of the degenerate torsional levels, $p_{v1,v1}$ is different from zero, so that the structure of the rotational levels here (called E rotational levels or E levels) differs from that of a rigid rotor. The linear term in P_z in the Hamiltonian has the effect of removing the K degeneracy characteristic of the rigid symmetric rotors. Furthermore, since the second-order perturbation term $w_{v\sigma}^{(2)}$ in Eq. (3-13b) is different for the A and E levels, each torsional sublevel will have a different effective rotational constant, $C_{v\sigma}$. The rotational levels with $\sigma=\pm1$ are doubly degenerate with respect to σ but nondegenerate with respect to K, and vice versa for the levels with $\sigma=0$. Hence, all the energy levels are doubly degenerate with the exception of the ones for which $\sigma=K=0$. Here the spatial M degeneracy has been disregarded. The three pairs of doubly degenerate levels $\sigma=0$, $\pm K$; $\sigma=+1$, $+K$; $\sigma=-1$, $-K$; and $\sigma=-1$, $+K$; $\sigma=+1$, $-K$ are shown in Fig. 4.

Physically, the presence of two sets of rotational levels with $\sigma=0$ and ±1 may be understood from a semiclassical point of view. For a high barrier the in-

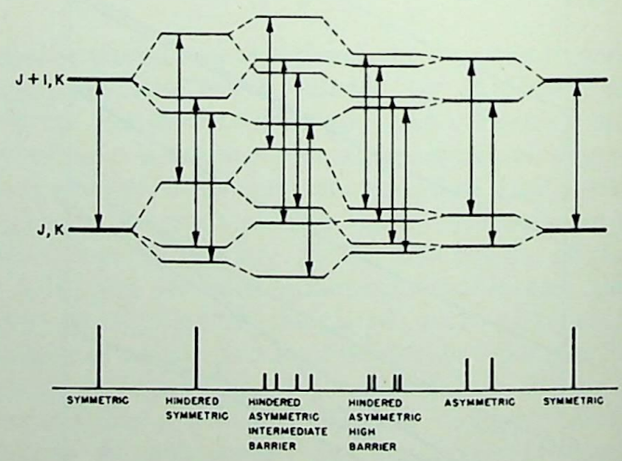


FIG. 4. Correlation of rotational energy levels. The rotational energy levels and rotational parallel ($\Delta K=0$) transitions for various conditions of asymmetry and barrier height. At either end the symmetric rotor energy levels and transition are shown. Introduction of hindered internal rotation splits each energy level in three levels for a threefold potential barrier. The three transitions however are coincident, giving only one line. Asymmetry splits the A level (K doubling), but only shifts the E levels. Two weaker lines on each side of the pattern are possible transitions between the E levels, but these are not shown in the figure. If the barrier height is raised, the energy levels tend toward the asymmetry doublets, and two doublets are observed in the spectrum. Finally, if the asymmetry is removed we are back to the symmetric rotor. The thickness of the lines for each energy level is approximately in agreement with the degeneracy of that level.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

ternal motion can be pictured as a back-and-forth oscillation about the equilibrium configuration of the molecule. The over-all rotation of the molecule along with this mode of internal torsion gives rise to the set of the A levels. The E levels can be associated with a different mode of internal motion in which the molecule passes from one equilibrium configuration to another in a circulatory manner through the tunneling effect. The double degeneracy of the E torsional levels is related to the two senses (counterclockwise and clockwise) of tunneling internal rotation. One may then expect the E levels and wave functions to have some free rotational character.

As the dipole moment of the molecule is independent of α, the selection rules for the rotational transitions are $\Delta J=\pm 1, \Delta K=0$, and $\Delta\sigma=0$ (see Sec. 6). Equation (3-13) shows that the frequencies of the rotational lines are not changed by the internal rotation. Both the A and E levels give rise to the same spectral lines and the three transitions shown in Fig. 4 are all coincident. Thus the effects of the internal rotation in the symmetric molecules are not detectable in the rotational spectra associated with the ground torsional state. This result can also be understood from the classical description by noticing that the rotation of a symmetric top about its figure axis causes no change in the orientation of the dipole moment and thus no dipole interaction with the radiation. When internal rotation is present in a symmetric top molecule, only the angular velocities of the two groups along the figure axis are affected.[67] In the excited torsional states, however, the frequencies of rotational transitions are appreciably altered by the internal rotation through the interaction with the other molecular vibrations.[42] The shifts of the frequencies of these lines then serve as a means of determining the barrier heights for the symmetric top molecules. This method is more fully discussed in Sec. 7.

5. Reduction to a rigid rotor at a very high barrier.— At the limit of a very high barrier, the torsional functions can be taken as Eq. (3-11). Furthermore, the overlapping between $H_v^{(1)}$, $H_v^{(2)}$, and $H_v^{(3)}$ can be neglected, i.e., the diagonal elements of p (i.e., $w_{v\sigma}^{(1)}$) approach zero. At first sight one might conclude that the effect of H_{IR} vanishes for infinitely high barriers and the Hamiltonian is then given by H_R. This, however, is not correct because the rotational constant of P_z^2 in H_R is not the rotational constant of the rigid molecule. The reason for this is that the off-diagonal elements of p increase with the barrier height so that the second-order term

$$\sum_{v'}{}'(b_{v\sigma}-b_{v'\sigma})^{-1}|p_{v\sigma,v'\sigma}|^2$$

remains finite in such a way that $w_v^{(2)}$ tends to zero. This latter fact would be expected since Eq. (3-13b) is in the form of a rigid molecule with the actual rotation constant. With Eq. (3-11) it can be shown that the only nonvanishing matrix elements are

$$p_{v,\sigma;v+1,\sigma}=\frac{1}{\sqrt{2}i}\left(\frac{9V_3}{4F}\right)^{\frac{1}{4}}(v+\tfrac{1}{2})^{\frac{1}{2}}$$
$$p_{v,\sigma;v-1,\sigma}=-\frac{1}{\sqrt{2}i}\left(\frac{9V_3}{4F}\right)^{\frac{1}{4}}v^{\frac{1}{2}}. \tag{3-14}$$

Furthermore, the energy of a torsional harmonic oscillator was given by Eqs. (3-4) and (3-10) as

$$E_v=3(v+\tfrac{1}{2})(V_3F)^{\frac{1}{2}}=(9/4)Fb_v.$$

When this equation and Eq. (3-14) are substituted into Eq. (3-13) the second-order term $w_{v\sigma}^{(2)}$ becomes zero. Therefore, the effective rotational constant $C_{v\sigma}$ becomes the rigid rotational constant C.

B. Internal Axis Method (IAM)

The essential feature of the IAM is the application of the Nielsen transformation to remove the coupling term between the angular momentum for over-all rotation and that associated with the internal rotation. Therefore, external and internal rotation become separable for the case of symmetric molecules. The IAM, however, has the disadvantage that the torsional functions are difficult to determine and are not conveniently tabulated.

For the PAM the coupling term H_{IR} (which is not small) is reduced by the Van Vleck transformation to a power series in K whose coefficients vanish at the limit of an infinite barrier. On the other hand, the coupling term H_{IR} is removed in the IAM by the Nielsen transformation, Eq. (2-8). The difference between these two transformations is that the Van Vleck is barrier dependent while the Nielsen is barrier independent.

The Hamiltonian was given in Eq. (2-20) as

$$H=A_a(P_a'^2+P_b'^2)+C_cP_c'^2+Fp'^2+V(\alpha),$$

where

$$A_a=A_x=\hbar^2/2I_a,$$
$$C_c=C=\hbar^2/2I_z(\neq C_z).$$

The a, b, c axes are the same as the principal axes for symmetric molecules, and the differential operators for P_a', P_b', P_c', and p' are given in Sec. 2 II. C2(b). Since the P''s do not contain α, the wave function can be written as a product of the symmetric rotor wave functions and the torsional wave function:

$$\psi=S_{JKM}(\theta,\varphi)e^{iKx'}M_{Kv\sigma}(\alpha'). \tag{3-15}$$

The symmetric rotor functions are the eigenfunctions of the first three terms of the Hamiltonian Eq. (2–20) and $M_{Kv\sigma}(\alpha')$ represent the eigenfunctions of the equation

$$\left[-F\frac{d^2}{d\alpha'^2}+V(\alpha')\right]M(\alpha')=EM(\alpha'), \tag{3-16}$$

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

with the appropriate boundary conditions. The symmetric top wave function of Eq. (3-15) contains χ' instead of χ as a variable. The angle χ' was defined in Eq. (2-15).

The boundary condition for Eq. (3-16) must now be considered. Obviously, the wave function must be periodic in 2π with respect to χ_1 and χ_2 (the Eulerian angles of the two parts), i.e.,

$$\psi(\theta,\varphi,\chi_1,\chi_2)=\psi(\theta,\varphi,\chi_1+2\pi n_1,\chi_2+2\pi n_2), \quad (3\text{-}17)$$

where n_1 and n_2 are integers. In terms of χ' and α', the wave function ψ becomes

$$\psi(\theta,\varphi,\chi',\alpha')=\psi\{\theta,\varphi,\chi'+(2\pi/I_c)[I_\alpha n_2+(I_c-I_\alpha)n_1], \alpha'+2\pi(n_2-n_1)\}. \quad (3\text{-}18)$$

In order to satisfy this condition use is made of the fact that according to Floquet's theorem, a particular (nonperiodic) solution of the Mathieu equation can be written as

$$M(\alpha)=e^{i\alpha f}P(\alpha), \quad (3\text{-}19)$$

where $P(\alpha)$ is periodic in 2π and f is a real constant which is chosen so that Eq. (3-19) has the periodicity as demanded by the physical situation. For the present case one has

$$(K/I_c)[I_\alpha n_2+(I_c-I_\alpha)n_1]+f(n_2-n_1)=n \text{ (an integer)}. \quad (3\text{-}20)$$

This can be satisfied by setting f equal to either $-\rho K$ [i.e., $-(I_\alpha K/I_c)$] or $(1-\rho)K$ [i.e., $+(I_c-I_\alpha/I_c)K$]. In the first of the series of papers on the internal rotation in methyl alcohol, Koehler and Dennison chose $(1-\rho)K$. [Also their internal angle χ is equal to $-\alpha$ which is used here (see Appendix 4).] However, it turns out, for most cases, to be convenient to take $f=-\rho K$. The resulting wave equation is then

$$\psi=S_{JKM}(\theta,\varphi)e^{iK\chi'}e^{-i\rho K\alpha'}P(\alpha'). \quad (3\text{-}21)$$

Equation (3-21) shows that the internal and external motion are separable inasmuch as the wave function can be expressed as a product of a function of the three external Eulerian angles and a function of α ($\alpha'=\alpha$) alone. The fact that the quantum number K appears in the torsional part of the wave function does not signify a physical coupling between the over-all rotation and the internal motion in the sense that the torsion of one group of the molecule about the other does not affect the motion of the "internal rotation axes" (see Sec. 2 II. A1). On the other hand, the reader may recall that in the treatment of the symmetric molecules by PAM there is a coupling term H_{IR} between the external and internal motion. In the PAM the set of moving axes is rigidly attached to one group of the molecule; while in the IAM the internal rotation axes, which are used as the coordinate system, are moving with respect to both of the groups. Since only the rotation of the internal rotation axes is unaltered by the internal motion, the coordinate frame used in PAM also oscillates at the frequency of the internal torsion.

1. Rotational equation.—The first three terms of the Hamiltonian Eq. (2-20) are the same as a rigid symmetric rotor molecule with energies as

$$H_R=A(P_a^2+P_b^2)+CP_c^2$$
$$E_{JKM}=AJ(J+1)+(C-A)K^2, \quad (3\text{-}22)$$

where

$$A=\hbar^2/2I_a \quad \text{and} \quad C=\hbar^2/2I_c.$$

2. Torsional function.—Although the torsional equations for the PAM and the IAM are identical, the torsional functions for these two cases are not the same because of the different boundary conditions. As explained above, the torsional function of IAM may be written as

$$M(\alpha)=e^{-i\rho K\alpha}P(\alpha). \quad (3\text{-}23)$$

The differential equation for $P(\alpha)$ is obtained by substituting Eq. (3-23) into Eq. (3-16) with the result

$$-F(d^2P/d\alpha^2)+2iF\rho K(dP/d\alpha)+\tfrac{1}{2}V_3(1-\cos 3\alpha)P=(E-F\rho^2K^2)P, \quad (3\text{-}24)$$

where

$$F\rho=\frac{\hbar^2}{2(I_c-I_\alpha)} \quad \text{and} \quad F\rho^2=\frac{\hbar^2}{2}\frac{I_\alpha}{I_c(I_c-I_\alpha)}.$$

This equation [or Eq. (3-16)] can be solved in the most general way by the method of continued fraction which is discussed in Appendix 2.

(*a*) *Properties of the torsional levels:* Since the solutions of Eq. (3-23) are different for different K, the value of K should be included as a labeling index for $P(\alpha)$ and E. Furthermore, as α appears explicitly in Eq. (3-16) only as $\cos 3\alpha$, an eigenfunction $M(\alpha)$ can be classified into one of the three general types corresponding to $\sigma=0,\pm1$††† in the expression

$$P_{Kv\sigma}=\sum_k A_{3k+\sigma}^{Kv}e^{i(3k+\sigma)\alpha},$$
or
$$M_{Kv\sigma}=e^{-i\rho K\alpha}\sum_k A_{3k+\sigma}^{Kv}e^{i(3k+\sigma)\alpha}. \quad (3\text{-}25)$$

Analogous to the solution of the torsional equation in the PAM, for each given σ one has a family of the functions $P_{Kv\sigma}(\alpha)$ and these functions are again labeled by the index v. Unlike the $U(\alpha)$'s given in Eq. (3-9), the eigenfunctions $P_{K,v,1}(\alpha)$ here are not degenerate with $P_{K,v,-1}(\alpha)$ because the differential Eq. (3-24) for $P(\alpha)$ contains imaginary coefficients. On the other hand, since $P^*(\alpha)$ is a solution of the equation obtained by taking the complex conjugate of Eq. (3-24) (the same

††† The quantity σ is the same as Itoh's $-\mu$ and is related to τ used by Dennison *et al.*[8,20,32,48] by the expression $K+\tau=\sigma+1$ (mod 3).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

as the differential equation for $P(\alpha)$ with $-K$), we have

$$\begin{aligned} P_{K,v,1}(\alpha) &= P_{-K,v,-1}{}^*(\alpha), \\ P_{K,v,0}(\alpha) &= P_{-K,v,0}{}^*(\alpha), \\ M_{K,v,\sigma}(\alpha) &= M_{-K,v,-\sigma}{}^*(\alpha), \\ P_{0,v,\sigma} &= U_{v,\sigma}, \end{aligned} \tag{3-26}$$

and

$$\begin{aligned} E_{K,v,1} &= E_{-K,v,-1}, \\ E_{K,v,0} &= E_{-K,v,0}. \end{aligned}$$

At high barriers the arrangement of the eigenvalues ($E_{Kv\sigma}$) in the IAM is similar to that in the PAM, i.e., the levels are grouped in a number of clusters corresponding to different values of the torsional quantum number v. Within each torsional state there are three distinct levels (rather than two in contrast with the results of the PAM) characterized by $\sigma=0, 1, -1$, for $K\neq 0$. The three levels are shown in Fig. 4. When K is equal to zero, the levels with $\sigma=1$ and $\sigma=-1$ are degenerate. The result that the eigenvalues are different for the PAM and the IAM comes as no surprise for the torsional equation is actually defined differently in these two methods. Physically, this discrepancy arises from the fact that the external rotation in the IAM is not the same as that in the PAM. However, the composite energy levels for the combination of the external and internal rotation calculated by both methods are identical as one would expect. This is illustrated in Sec. I. B3.

(*b*) *Eigenvalues of the torsional equation:* The tabulated solutions of the Mathieu equation are not directly applicable to the torsional function in the IAM since these functions do not have the periodicity of 2π. The general method for obtaining the solution by the use of continued fraction is described in Appendix 2. The procedure for determining the eigenfunctions is in general complicated but at the very high barrier these functions can be approximated by the harmonic oscillator functions. In addition, the eigenvalues can be simply obtained by the following approximate method: when $\rho K-\sigma$ is replaced by $\rho K-\sigma+3$, the form of Eq. (3-25) is not altered, and therefore the eigenvalues remain unchanged. One may then regard the eigenvalues as periodic functions of $(2\pi/3)(\rho K-\sigma)$ and expand them in a Fourier series as

$$E_{Kv\sigma}=F\sum_n a_n{}^{(v)}\cos(2\pi n/3)(\rho K-\sigma). \tag{3-27}$$

If $\rho K-\sigma$ is equal to zero or a whole multiple of three, then Eq. (3-25) reduces to the solution of the Mathieu equation (with a periodicity of π) when the substitutions given in Eq. (3-4) are made, i.e.,

$$E(\rho K-\sigma=0)\rightarrow b(\pi).$$

Here b and E differ by the scale factor given by Eq. (3-4). Similarly, one can show that

$$\begin{aligned} E(\rho K-\sigma=\tfrac{3}{2}) &\rightarrow b(2\pi), \\ E(\rho K-\sigma=1) &\rightarrow b(3\pi), \end{aligned}$$

and more generally

$$E(\rho K-\sigma=x)\rightarrow b(3\pi/x).$$

Up to this time only the low integral values of $3/x$ have been tabulated (see Appendix 3).

If the series in Eq. (3-27) converges so rapidly that E may be approximated by the first two terms in the Fourier expansion, i.e.,

$$E=F[a_0+a_1\cos(2\pi/3)(\rho K-\sigma)], \tag{3-28}$$

then the coefficients a_0 and a_1 may be determined from the tabulated values[104] of $b(\pi)$ and $b(2\pi)$. When this is done, Eq. (3-28) can be used to calculate the approximate values of $b(N\pi)$. To this approximation the equation for $b(N\pi)$ is

$$b(N\pi)=\frac{1}{2}\left(1+\cos\frac{2\pi}{N}\right)b(\pi) + \frac{1}{2}\left(1-\cos\frac{2\pi}{N}\right)b(2\pi), \tag{3-29}$$

such that any Mathieu eigenvalue can be approximated[23] from $b(\pi)$ and $b(2\pi)$ provided that the Fourier coefficient a_1 is small. If the coefficient a_1 is not small then $b(3\pi)$, $b(4\pi)$, etc., can be used to obtain the coefficients a_2 and a_3 and a better value of $b(N\pi)$ (see Table III). Therefore, the Mathieu eigenvalues can be readily approximated for any periodicity.

(*c*) *Harmonic oscillator approximation:* As in the case of the PAM, the torsional function can be approximated by harmonic oscillator wave functions at very high barriers. Since the classical amplitude of the internal torsion approaches zero at the limit of very high barriers, the factor $e^{-i\rho K\alpha}$ in the torsional function $M(\alpha)$ may be replaced by unity and $M_{Kv\sigma}(\alpha)\rightarrow P_{Kv\sigma}(\alpha)$. Under such circumstances $M(\alpha)$ can be approximated by Eq. (3-11). The energy levels of Eq. (3-16) are then given by Eq. (3-10).

3. Energy levels and spectra.—The total energy of a symmetric top exhibiting internal rotation is equal to the sum of the energy associated with the over-all rotation E_{JKM} and the torsional energy $E_{Kv\sigma}$, or

$$E(JKMv\sigma)=E_{JKM}+E_{Kv\sigma}. \tag{3-30}$$

Since the K degeneracy remains in the rigid symmetric rotor energy, from Eq. (3-26) it may be seen that

$$E(J,K,M,v,\sigma)=E(J,-K,M,v,-\sigma).$$

Therefore, all the energy levels are doubly degenerate except the $K=0, \sigma=0$ levels—a result which was also found by the PAM. Now substituting Eqs. (3-22) and (3-27) into Eq. (3-30), we obtain

$$E(JKMv\sigma)=AJ(J+1)+(C-A)K^2 + F\sum_n a_n{}^{(v)}\cos\frac{2\pi n}{3}(\rho K-\sigma). \tag{3-31a}$$

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

The last term in this equation represents the effect of the internal torsion on the rotational energy levels. If ρK is reasonably small and the Fourier series converges rapidly, the cosine terms in the last member of Eq. (3-31) can be expanded in the following manner:

$$\sum a_n^{(v)} \cos\frac{2\pi n}{3}(\rho K-\sigma)$$

$$=\sum_n a_n^{(v)} \cos\frac{2\pi\sigma}{3}\left[1-\frac{1}{2}\left(\frac{2\pi n}{3}\rho K\right)^2+\cdots\right]$$

$$+\sum a_n^{(v)} \sin\frac{2\pi\sigma}{3}\left[\frac{2\pi n}{3}\rho K+\cdots\right]. \quad (3\text{-}31b)$$

For $\sigma=0$ the K^2 term from Eq. (3-31b) can be incorporated into the term $(C-A)K^2$ in Eq. (3-31a), and therefore it has the effect of modifying the rotational constant. The rotational levels for this case have the same pattern as a rigid symmetric rotor—pseudo-rigid rotor level. In the case of the degenerate torsional levels ($\sigma=\pm1$), the expansion in Eq. (3-31b) introduces a term linear in K, into the energy equation. This causes a splitting of the K degeneracy of the rigid symmetric rotors. The conclusions obtained here are therefore in complete agreement with those derived from the PAM. When ρK becomes larger, the $(\rho K)^3$ and $(\rho K)^4$ terms should be included in the expansion of $\cos(2\pi/3)(\rho K-\sigma)$. In terms of the PAM the higher-order Van Vleck perturbation treatment must be applied to the Hamiltonian in order to obtain accurate results.

By following the selection rules $\Delta J=\pm1$, $\Delta K=0$, and $\Delta\sigma=0$ one finds again that the rotational frequencies depend only on the rotational constant A and the quantum number J, and are not affected by the internal torsion. Therefore, only one line is observed as shown in Fig. 4 for this model.

C. Comparison of the Two Methods

The main difference between the two methods is that the IAM gives a relatively simple Hamiltonian but more complicated torsional functions, whereas the PAM gives a more complex Hamiltonian but simpler torsional functions. The chief difficulty in the IAM is associated with the determination of the torsional eigenfunctions and eigenvalues. On the other hand, the coupling term pP_z in the PAM is not a small perturbation, so fourth-order perturbation calculation is sometimes necessary to obtain accurate results for the energy levels.

A comparison of Eqs. (3-13b) and (3-31a) shows that the effects of the internal torsion on the rotational energy levels are expressed by a power series in the PAM and a Fourier series in the IAM. Generally, calculation of the first- and second-order terms in the power series ($w_{v,\sigma}^{(1)}$ and $w_{v,\sigma}^{(2)}$) is comparatively simple, but the higher-order terms are more difficult to evaluate. Herschbach[23] has suggested that the tabulated eigenvalues of the Mathieu equation and the first two perturbation terms in PAM be used to determine the Fourier coefficients in the IAM which are otherwise tedious to calculate by the method of continuous fraction. Since the Fourier series in IAM converges much faster than the power series in Eq. (3-13b), the former may be used to obtain the rotational energy levels and also the higher-order perturbation terms of the PAM. This technique is really very interesting for we have solved the problem approximately by the PAM where the boundary conditions are simple and have then converted this to the more exact solution of the IAM.

II. Asymmetric Molecules

A. *Principal Axis Method* (PAM)

The Hamiltonian given in Eq. (2-37) can be grouped in the following manner:

$$H_0=\tfrac{1}{2}(A_x+B_x)(P_x^2+P_y^2)+C_zP_z^2+Fp^2+V(\alpha), \quad (3\text{-}32a)$$

$$H_1=-2\sum_i Q_iP_ip, \quad (3\text{-}32b)$$

and

$$H_2=\tfrac{1}{2}(A_x-B_x)(P_x^2-P_y^2)$$
$$+\tfrac{1}{2}\sum_{i\neq j} D_{ij}(P_iP_j+P_jP_i). \quad (3\text{-}32c)$$

The energy matrix is now constructed with the basis functions

$$S_{JKM}(\theta,\varphi)e_i^{Kx}U_{v\sigma}(\alpha).$$

In this representation H_0 is completely diagonal; H_2 is diagonal in v, but not in K; and H_1 has diagonal and off-diagonal matrix elements in K and v. As in the case of symmetric molecules, the Van Vleck transformation is applied to remove the nondiagonal elements in v so that the new energy matrix assumes the block structure shown in Fig. 3. The effective rotational Hamiltonian, correct to second order, for the vth torsional state is‡‡‡

$$H_{v\sigma}=A_{v\sigma}P_x^2+B_{v\sigma}P_y^2+C_{v\sigma}P_z^2+E_{v\sigma}$$
$$+\tfrac{1}{2}\sum_{i\neq j} D_{ij}w_{v\sigma}^{(2)}(P_iP_j+P_jP_i)+\sum_i Q_iP_iw_{v\sigma}^{(1)}, \quad (3\text{-}33)$$

where

$$A_{v\sigma}=\frac{h^2}{2I_x}\left[1+\frac{\lambda_x^2I_\alpha}{rI_x}w_{v\sigma}^{(2)}\right],$$

$$w_{v\sigma}^{(1)}=-2p_{v\sigma,v\sigma},$$

$$w_{v\sigma}^{(2)}=1+\frac{16}{9}\sum_{v'}{}'\frac{|p_{v\sigma,v'\sigma}|^2}{b_{v\sigma}-b_{v'\sigma}},$$

‡‡‡ Here we have assumed that the torsional levels of different v are much more widely spaced compared to the rotational levels. When this is not the case, the $(K|K')$ elements, introduced by H_1 to the diagonal v blocks through the Van Vleck transformation, have a very complicated form.[25b]

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

and similarly for B_v and C_v with the appropriate permutations of x, y, and z.

Alternatively the Hamiltonian given in Eq. (2-38) could be used for the perturbation treatment. The Van Vleck transformation now gives the following[23]:

$$H_{v\sigma}=\frac{h^2}{2}\sum_g\frac{P_g^2}{I_g}+F\sum_n w_{v\sigma}^{(n)}\Pi^n, \qquad (3\text{-}34)$$

where

$$\Pi=\varrho\cdot\mathbf{P},$$

and the $w_{v\sigma}^{(n)}$'s are the same perturbation terms given in Eq. (3-13b). Naturally, Eqs. (3-33) and (3-34) are identical. Equation (3-33) shows the perturbation calculation more clearly and shows that the rotational constants are modified by the internal rotation. Equation (3-34) shows the similarity to the symmetric molecule formulas and leads more naturally, as we will see later, to higher-order perturbation corrections.

1. Energy levels and spectra.—From Eq. (3-33) the rotational constants for the over-all rotation of the molecule are modified by the internal rotation. Also the orientation of the "effective" principal axes is changed by the cross terms $D_{ij}w_{v\sigma}^{(2)}(P_iP_j+P_jP_i)$. Both of these changes depend on the torsional state v and σ. As in the case of the symmetric molecules, there exist two sets of rotational levels for each torsional state. The Hamiltonian for the pseudo-rigid rotor levels ($\sigma=0$) to second order is given by Eq. (3-35),

$$H_{v0}=A_{v0}P_x^2+B_{v0}P_y^2+C_{v0}P_z^2+\sum_{i\neq j}\tfrac{1}{2}D_{ij}w_{v0}^{(2)}(P_iP_j+P_jP_i), \qquad (3\text{-}35)$$

where the rotational constants A_{v0}, etc., differ from the respective constants of a rigid molecule by terms proportional to $w_{v0}^{(2)}$. Also the cross terms between the P's contain $w_{v0}^{(2)}$. As mentioned before, the quantity $w_{v0}^{(2)}$ decreases rapidly with increasing value of the reduced barrier height s [defined in Eq. (3-4)] and approaches zero for very high barrier.

For the E levels the Hamiltonian is also given by Eq. (3-33). Because of the linear terms in P_x, P_y, and P_z the spacings of the E levels are different even qualitatively from those of a pseudo-rigid rotor. However, in some cases, especially for high barriers and low K energy levels, when the asymmetry splitting of the K doublet is not too small, the effect of the terms containing $p_{v\sigma,v\sigma}$ is negligible.§§§ The effective rotational constants $A_{v,\pm1}$, etc., are, however, appreciably different from those for the A states as well as those for the rigid molecule. Under such circumstances both sets of the A and E levels are similar in structure but have different effective rotational constants.

§§§ It should be noted that the effect of the linear terms is negligible only when it connects no near degenerate levels. Also if the quantity ρK becomes large, higher-order perturbation terms are needed [see Eq. (3-34) and Sec. 3].

The selection rules are identical to those for a rigid asymmetric top with the additional restriction of $\Delta\sigma=0$ since the dipole moment is independent of the angle α. The most intense spectrum arises from the molecules in the lowest torsional state, i.e., $v=0$. For each rotational transition one finds two lines; one originates from the rotational levels associated with the nondegenerate torsional levels and the other one from the degenerate E levels. The essential feature of the spectrum is that doublets occur instead of a single line as in the case of a rigid molecule. This is illustrated in Fig. 4. The lines belonging to the A species follow a pseudo-rigid spectrum to the second order and the E lines may or may not be fitted to a pseudo-rigid Hamiltonian depending on the height of the barrier and the asymmetry of the molecule. The two members of the doublet are of more or less the same intensity; the ratio of the intensity is governed by the relative nuclear statistical weight assigned to the A and E levels (see Sec. 6). The separation of the two lines of a doublet depends very critically on the height of the barrier. It is from these splittings that we determine the barrier height. In the next section we discuss the calculation of the doublet splittings. The methods of evaluation of the barrier height from the measured values of the splittings are presented in Sec. 5.

2. Calculation of energy and doublet splittings.—The matrix elements of $H_{v\sigma}$ in Eq. (3-33)[101] are

$$\begin{aligned}
(K|H_{v\sigma}|K)&=\tfrac{1}{2}(A_{v\sigma}+B_{v\sigma})J(J+1)\\
&\quad+[C_{v\sigma}-\tfrac{1}{2}(A_{v\sigma}+B_{v\sigma})]K^2+Q_zw_{v\sigma}^{(1)}K,\\
(K|H_{v\sigma}|K\pm1)&=\tfrac{1}{2}[(D_{yz}\pm iD_{xz})w_{v\sigma}^{(2)}(2K\pm1)\\
&\quad+(Q_y\pm iQ_x)w_{v\sigma}^{(1)}][J(J+1)-K(K\pm1)]^{\frac{1}{2}},\\
(K|H_{v\sigma}|K\pm2)&=-\tfrac{1}{4}[A_{v\sigma}-B_{v\sigma}\pm2iD_{xy}w_{v\sigma}^{(2)}]\\
&\quad\times[J(J+1)-K(K\pm1)]^{\frac{1}{2}}\\
&\quad\times[J(J+1)-(K\pm1)(K\pm2)]^{\frac{1}{2}}. \qquad (3\text{-}36)
\end{aligned}$$

The indices J, M have been deleted since $H_{v\sigma}$ is diagonal in these quantum numbers, and the constant term $E_{v\sigma}$ has been dropped. Here, in addition to the $(K|K\pm2)$ contribution from the usual asymmetry of H_2, there are off-diagonal elements from the cross terms in $D_{ij}w_{v\sigma}^{(2)}$. Also H_1 has contributed both a diagonal term in $Q_zw_{v\sigma}^{(1)}$ and off-diagonal terms in $Q_xw_{v\sigma}^{(1)}$ and $Q_yw_{v\sigma}^{(1)}$.

The Hamiltonian $H_{v\sigma}$ in Eq. (3-33) may be simplified by performing a rotation of coordinate axes to eliminate the cross terms between the components of the total angular momentum P. Referred to this new coordinate axes system which is the "effective principal axes" system for a torsional level, the effective Hamiltonian becomes

$$H_{v\sigma}'=A_{v\sigma}'P_x'^2+B_{v\sigma}'P_y'^2+C_{v\sigma}'P_z'^2+\sum_i Q_i'P'w_{v\sigma}^{(1)}. \qquad (3\text{-}37)$$

The orientation between the two sets of axes can be determined either by diagonalizing the 3×3 matrix in

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

TABLE I. Low J energy levels.[a]

E_{JK}

$E_{00}=0$

$E_{10}=A_{v\sigma}'+B_{v\sigma}'-(g^{+}+g^{-})$

$E_{11}=\frac{1}{2}(A_{v\sigma}'+B_{v\sigma}')+C_{v\sigma}'+\frac{1}{2}(g^{+}+g^{-})$
$\pm[(Q_x'w_v^{(1)})^2+\frac{1}{4}(A_{v\sigma}'-B_{v\sigma}')^2-(A_{v\sigma}'-B_{v\sigma}')(g_y-g_x)]^{\frac{1}{2}}$

$E_{20}=W_r(A_{v\sigma}'B_{v\sigma}'C_{v\sigma}'J=2, K=0)-3(g^{+}+g^{-})$
$+0[(A-B)(2C-A-B)^{-1}g]$

$E_{21}=(5/2)(A_{v\sigma}'+B_{v\sigma}')+C_{v\sigma}'+3(g^{+}+g^{-})/2-(k^{+}+k^{-})$
$\pm[(Q_z'w_v^{(1)})^2+(9/4)(A_{v\sigma}'-B_{v\sigma}')^2-9(A_{v\sigma}'-B_{v\sigma}')(g_y-g_x)]^{\frac{1}{2}}$

where

$g^{\pm}=(Q_x'^2+Q_y'^2)[w_{v\sigma}^{(1)}]^2[2C_{v\sigma}'-A_{v\sigma}'-B_{v\sigma}'\pm 2Q_z'w_{v\sigma}^{(1)}]^{-1}$

$g_i=Q_i'^2[w_{v\sigma}^{(1)}]^2[2C_{v\sigma}'-A_{v\sigma}'-B_{v\sigma}']^{-1}(i=x, y)$

$g_x+g_y\approx\frac{1}{2}(g^{+}+g^{-})$

$k^{\pm}=2(Q_x'^2+Q_y'^2)[w_{v\sigma}^{(1)}]^2[6C_{v\sigma}'-3A_{v\sigma}'-3B_{v\sigma}'\pm 2Q_z'w_{v\sigma}^{(1)}]^{-1}$

[a] The $A_{v\sigma}'$, $B_{v\sigma}'$, and $C_{v\sigma}'$ are not necessarily identical with the A, B, and C of King *et al.*,[41] and the quantum number K is no longer a good quantum number, but it is still used for labeling. In the expression for the E_{20} energy W_r is the rigid rotor energy with the rotational constants A_v', B_v', and C_v'.

the usual manner or by applying successive 2×2 rotations. The relations between the new coefficients $A_{v\sigma}'$, etc., in Eq. (3-37) and the old ones in Eq. (3-33) can be set down by noting that A, B, C, D_{xy}, etc., transform like the six components of a symmetric second-rank tensor and Q_x, Q_y, Q_z, transform like the components of a vector. Very often the change in the effective rotational constants introduced by such a rotation is negligible. For instance, in the case of acetaldehyde (CH_3CHO) the cross term P_iP_j is removed by rotating the body axes through an angle of approximately one degree.

For the nondegenerate A levels the last term of Eq. (3-37) vanishes. The rotational energies are just those for a pseudo-rigid rotor with slightly corrected rotational constants and can be calculated by well-known methods or obtained from published tables.[41,97,108] In the case of the E levels the energy matrices for both Eq. (3-33) and Eq. (3-37) contain a term linear in K on the diagonal and also $(K|K\pm1)$ elements. The two kinds of factoring available for rigid asymmetric tops are spoiled, and for a given J, a $(2J+1)(2J+1)$ determinantal equation must be solved. Except for low J, the straightforward diagonalization of this matrix is very difficult. The procedure of diagonalization may be somewhat simplified by applying a Van Vleck transformation to reduce the $(K|K\pm1)$ elements. Except for the energy levels with small asymmetry splitting, after a second-order Van Vleck transformation the new $(K|K\pm1)$ elements are usually such that their contributions to the diagonal terms are negligible for low J levels. The approximate formulas for some of the energy levels are given in Table I. Also approximate formulas for the splitting of the A and E levels (not the lines) can be derived; these are given in Table II.

When the barrier is very high the value of $w_{v\sigma}^{(1)}$ becomes so small that the E levels for low K also conform to the pattern of a rigid rotor with effective rotational constants different from those belonging to the A levels. The energy splitting between the A and E levels, ΔW, is simply equal to the difference between the "rigid rotor energy" of the two types of levels, W_A, and W_E, as

$$\Delta W=W_A-W_E=(\partial W/\partial A)\Delta A+(\partial W/\partial B)\Delta B+(\partial W/\partial C)\Delta C. \quad (3\text{-}38)$$

If the rigid rotor energy is written in the usual form[41]

$$W=\tfrac{1}{2}(A+C)J(J+1)+\tfrac{1}{2}(A-C)E(\kappa),\|\|\| \quad (3\text{-}39)$$

then it follows that

$$\begin{aligned}\partial W/\partial A&=\tfrac{1}{2}J(J+1)+\tfrac{1}{2}E(\kappa)-\tfrac{1}{2}(\kappa+1)[\partial E(\kappa)/\partial\kappa],\\ \partial W/\partial B&=\partial E(\kappa)/\partial\kappa,\\ \partial W/\partial C&=\tfrac{1}{2}J(J+1)-\tfrac{1}{2}E(\kappa)+\tfrac{1}{2}(\kappa-1)[\partial E(\kappa)/\partial\kappa].\end{aligned} \quad (3\text{-}40)$$

The values of $[\partial E(\kappa)/\partial\kappa]$ can be estimated from the tabulated eigenvalues of rigid rotors.[41,108] By substituting the results of Eq. (3-40) into Eq. (3-38) the energy separation between the A and E levels is obtained.

3. Higher-order effects.—(a) Third- and fourth-order perturbation terms[23,25b]: Up to this point the discussion of the PAM represents the second-order theory in the Van Vleck perturbation treatment. When the barrier is not sufficiently high and/or the quantity ρK is large (as in the case of CH_3COOH),[105] the third- and fourth-order perturbation terms should be taken into consideration. The form of the effective Hamiltonian given in Eq. (3-34) is very convenient for the calculation of the higher-order terms. Also, the perturbation coefficients are the same as those given in Eq. (3-13b) for a symmetric molecule. The higher-order perturbation corrections can be calculated directly from the matrix elements of p or by the labor-saving method given in Sec. I. C. Herschbach[23,25b] has tabulated a number of these terms. The readers are directed to Sec. 5 for more detail of this method. With the higher-order terms,

TABLE II. Low J energy splitting $(A-E)$.

$\Delta E_{00}=0$

$\Delta E_{10}=\Delta A+\Delta B+g^{+}+g^{-}$

$\Delta E_{11}=\Delta C+\frac{1}{2}(\Delta A+\Delta B)-\frac{1}{2}(g^{+}+g^{-})$
$\pm\left[\frac{1}{2}(\Delta A-\Delta B)+(g_y-g_x)-\frac{Q_z^2[w_{v\sigma}^{(1)}]^2}{A-B}\right]$

$\Delta E_{21}=\Delta C+(5/2)(\Delta A+\Delta B)-(3/2)(g^{+}+g^{-})+(k^{+}+k^{-})$
$\pm\left[(3/2)(\Delta A-\Delta B)+3(g_y-g_x)-\frac{Q_z^2[w_{v\sigma}^{(1)}]^2}{3(A-B)}\right]$

where

$$|Q_zw_{v\sigma}^{(1)}|<|(A-B)|$$

$\|\|\|$ In this equation the rotation constants A, B, C, are such that $A\geq B\geq C$ and they are not necessarily identical to $A_{v\sigma}$, $B_{v\sigma}$, and $C_{v\sigma}$.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

matrix elements of the form $(K|K\pm3)$ and $(K|K\pm4)$ are introduced and these can be treated in a manner similar to Kivelson and Wilson's treatment[47] of the centrifugal distortion problem.

(*b*) *Sixfold potential energy term*, V_6[25b,38]: So far we have assumed that the potential function is sinusoidal in α without rigorous justification. If the expansion of the potential function in Eq. (2-1b) converges rapidly, the V_6 term can be treated as a perturbation. When this is done, the V_6 contribution merely changes the values of all the perturbation coefficients $w_{v\sigma}^{(n)}$ in such a way that its contribution cannot be separated from that of V_3 without data from more than one torsional state. In order to evaluate the V_6 coefficient, it is necessary then to study the internal rotational effects in more than one torsional state.¶¶¶ This V_6 term should be generally smaller than the threefold potential by a factor ranging from ten to a thousand.

B. Internal Axis Method (IAM)

Since the treatment by the IAM for the most general type of molecule (no symmetry aside from that of the CH_3 group) is algebraically very complicated, the discussion here is confined to those molecules which have a planar framework. Papers by Pitzer and Gwinn,[84] Hecht and Dennison,[20] and Burkhard[7] discuss the general type of asymmetric molecule by the IAM.

1. Hamiltonian and matrix elements.—In Sec. 2 three different forms of the Hamiltonian were derived. They were given in Eqs. (2-26), (2-29), and (2-31). We are now at liberty to use any one of these Hamiltonians to solve for the energy levels. Let us first consider the Hamiltonian (2-29).[20,31] This Hamiltonian can be divided as follows:

$$H_0=\tfrac{1}{2}(A_a+B_b'')(P_a''^2+P_b''^2)+C_c''P_c''^2+Fp'^2+V(\alpha),$$
$$H_1=\tfrac{1}{2}(A_a-B_b'')(P_a''^2-P_b''^2)+D_{bc}''(P_b''P_c''+P_c''P_b'').$$

P_a'', P_b'', and P_c'' are the components of the total angular momentum along the axes fixed in the framework of the molecule (rather than the internal axes which rotate with respect to both the framework and the top), i.e., the orientation of this axes system is described by the Eulerian angles θ, Φ, and χ (but not χ'). Consequently, we choose the eigenfunctions of H_0 as the basis functions for constructing the energy matrix. These functions are

$$\psi=S_{JKM}(\theta,\Phi)e^{iK\chi}Q(\alpha), \tag{3-41}$$

where $Q(\alpha)$ is the solution of the torsional equation

$$[Fp'^2+V(\alpha)]Q(\alpha)=EQ(\alpha). \tag{3-42}$$

Since P_c'' is diagonal in H_0, using Eq. (2-28) we can write

$$p'=p-\rho P_c''=(1/i)(\partial/\partial\alpha)_{\theta,\Phi,\chi}-\rho K, \tag{3-43}$$

which can then be substituted into Eq. (3-42) with the result

$$[F(-i(d/d\alpha)-\rho K)^2+V(\alpha)]Q(\alpha)=EQ(\alpha), \tag{3-44a}$$

or

$$[-F(d^2/d\alpha^2)+V(\alpha)]M(\alpha)=EM(\alpha), \tag{3-44b}$$

where $M(\alpha)$ is $e^{-i\rho K\alpha}Q(\alpha)$. Equations (3-44) are just the same as Eqs. (3-16) and (3-24) only with the scale change in ρ and F [see Eq. (2-30)]. Analogous to the function $P(\alpha)$ in Eq. (3-25) the eigenfunction $Q(\alpha)$ can be expanded in a Fourier series as

$$Q(\alpha)=\sum_k A_{3k+\sigma}{}^{Kv}e^{i(3m+\sigma)\alpha}. \tag{3-45}$$

Furthermore, it can be easily seen that

$$Q_{Kv\sigma}(\alpha)=Q_{-Kv-\sigma}{}^*(\alpha),$$
$$E_{Kv\sigma}=E_{-Kv-\sigma}. \tag{3-46}$$

Before deriving the matrix elements of H let us consider the Hamiltonian[20] given in Eq. (2-31). This Hamiltonian can be divided as follows:

$$H_0=\tfrac{1}{2}(A_a+B_b'')(P_a'^2+P_b'^2)+C_c''P_c'^2+Fp'^2+V(\alpha)$$
$$H_1=\tfrac{1}{2}(A_a-B_b'')$$
$$\times[(P_a'^2-P_b'^2)\cos2\rho\alpha-(P_a'P_b'+P_b'P_a')\sin2\rho\alpha]$$
$$+D_{bc}''[(p_b'p_c+P_c'P_b')\cos\rho\alpha$$
$$+(p_a'P_c'+P_c'P_a')\sin\rho\alpha]. \tag{3-47}$$

Here P_a', P_b', and P_c' are components of the total angular momentum along the internal rotational axes. The orientation of this axes system is described by the Eulerian angles $\theta\Phi\chi'$ where

$$\chi'=\chi+\rho\alpha,$$

[see Eq. (2-28)]. The basis functions again are chosen to diagonalize H_0. They are

$$\psi=S_{JKM}(\theta,\Phi)e^{iK\chi'}M(\alpha'), \tag{3-48}$$

where $M(\alpha')$ is the solution of the torsional equation (3-44b) which must satisfy the boundary conditions [see Eq. (3-20) Sec. I. B]. Furthermore, we can see that the basis functions given in Eqs. (3-41) and (3-48) are indeed identical as

$$e^{iK\chi}Q(\alpha)=e^{iK\chi'}e^{-i\rho K\alpha}Q(\alpha)=e^{iK\chi'}M(\alpha').$$

The energy matrix can then be constructed from the Hamiltonian Eq. (2-29) with Eq. (3-41) as the basis functions or from Eqs. (2-31) and (3-48). The results of these two schemes are identical.

The matrix elements of the Hamiltonian[20,31] are as follows:

$$(Kv\sigma|H|Kv\sigma)=\tfrac{1}{2}(A_a+B_b'')[J(J+1)-K^2]$$
$$+C_c''K^2+E_{Kv\sigma}, \tag{3-49a}$$

¶¶¶ Kilb *et al.*[38] calculated a V_6 term but their approximation is now realized to be inadequate because they ignored the third-order correction in the evaluation of the V_6 term from the splittings in the ground torsional states.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

$$(Kv\sigma|H|K+1, v'\sigma)=D_{bc}''(K+\tfrac{1}{2})\times[J(J+1)-K(K+1)]^{\frac{1}{2}}I_{K+1,v'\sigma}{}^{Kv\sigma}, \quad (3\text{-}49b)$$

$$(Kv\sigma|H|K+2, v'\sigma)=-\tfrac{1}{4}(A_a-B_b'')[J(J+1)-K(K+1)]^{\frac{1}{2}}\times[J(J+1)-(K+1)(K+2)]^{\frac{1}{2}}I_{K+2,v'\sigma}{}^{Kv\sigma}, \quad (3\text{-}49c)$$

where

$$I_{K'v'\sigma}{}^{Kv\sigma}=\int_0^{2\pi} Q_{Kv\sigma}{}^*(\alpha)Q_{K'v'\sigma}(\alpha)d\alpha.^{****} \quad (3\text{-}50)$$

If the Hamiltonian form given in Eq. (2-26) were to be used,[8,31] then the scale factors ρ and F would be those corresponding to the symmetric molecule, i.e., $\rho=I_\alpha/I_{cc}$ and $F=(\hbar^2/2)[I_{cc}/I_\alpha(I_{cc}-I_\alpha)]$. Unfortunately, there are more matrix elements (those linear in p') which needed to be calculated and generally this formulation has no advantage over the other two. As a result we do not discuss it further.

2. Integrals involving the torsional functions.[20]—In order to obtain the energy levels it is necessary first to compute the integrals given in Eq. (3-50). Evaluation of such integrals is generally quite complicated because of the difficulty involved in the determination of the torsional functions by the procedure of continued fractions. Fortunately, for the case of high barriers ($s\gtrsim 20$), one may use the harmonic oscillator approximation for the torsional functions and the integrals can be obtained in a much simpler way.

At the limit of infinite barrier, $I_{K'v'\sigma}{}^{Kv\sigma}$ approaches zero for $v\neq v'$ and unity for $v=v'$. It is convenient then to introduce the terms

$$\delta_{K+1,v\sigma}{}^{Kv\sigma}=1-\int_0^{2\pi} Q_{Kv\sigma}{}^*(\alpha)Q_{K+1,v\sigma}(\alpha)d\alpha,$$
$$\delta_{K+2,v\sigma}{}^{Kv\sigma}=1-\int_0^{2\pi} Q_{Kv\sigma}{}^*(\alpha)Q_{K+2,v\sigma}(\alpha)d\alpha. \quad (3\text{-}51)$$

By using the harmonic oscillator approximation for the torsional wave function Hecht and Dennison[20] obtained the following formulas for the ground torsional state $v=0$:

$$\delta_{K+1,0\sigma}{}^{K0\sigma}=f(\rho/3),$$
$$\delta_{K+2,0\sigma}{}^{K0\sigma}=f(2\rho/3), \quad (3\text{-}52)$$

where

$$f(x)=\frac{1}{\sqrt{s}}x^2+\frac{1}{2s}x^2(1-x^2)+\cdots.$$

Expressions for $I_{K'v'\sigma}{}^{Kv\sigma}(v\neq v')$ have also been derived by Hecht and Dennison from the harmonic oscillator approximation [see Eq. (28) of reference 20]. An important property of the integrals in Eq. (3-50) is their relative independence of the quantum numbers K and σ. Furthermore, unlike the case of PAM the matrix elements off-diagonal in v are usually very small and their contributions to the energy are rather unimportant.

3. Diagonalization of the energy matrix.—The diagonalization of those elements off-diagonal in v can be accomplished by a Van Vleck transformation. After this transformation the energy matrix takes on a block structure in which each block corresponds to a certain value of J with the dimensions $(2J+1)\times(2J+1)$. Both $(K|K\pm1)$ and $(K|K\pm2)$ elements are present in these submatrices because the coordinate axes are not principal axes. One then applies a transformation S $(=S_1S_2)$ which diagonalizes the limiting rigid rotor Hamiltonian. The off-diagonal elements of the transformed Hamiltonian now consist of only the terms containing the integrals $I_{K'v'\sigma}{}^{Kv\sigma}$ and $E_{Kv\sigma}$. Examples of the transformed Hamiltonian have been given by Hecht and Dennison for $J=1$.†††† The energy levels can then be obtained by the usual perturbation method (or a modified procedure may be recommended in case near degeneracy occurs). One may, of course, apply a transformation S' to H to bring it directly into the diagonal form. The advantage of using S instead of S' is that the former transformation matrix can be set down in a systematic fashion.

Let S be written as S_1S_2 where S_1 transforms the coordinate axes to the principal axes system and S_2 is the matrix which diagonalizes the asymmetric rigid-rotor Hamiltonian (referred to principal axes) in the symmetric top representation. The elements of S_2 can be determined from the asymmetric rotor energy levels by well-known techniques.[97] The elements of S_1 can be written directly from the representation of the rotational group as given by Wigner[117] as

$$\psi_{JK_p}=\sum_K D^J(\beta)_{K_pK}\psi_{JK}=\sum_K(S_1)_{K_pK}\psi_{JK}, \quad (3\text{-}53)$$

where K_p and K refer to the quantum number along the z principal axis and the internal axis, respectively, and

$$D^J(\beta)_{K_pK}=\sum(-1)^n \times\frac{[(J+K)!(J-K)!(J+K_p)!(J-K_p)!]^{\frac{1}{2}}}{(J-K_p-n)!(J+K-n)!n!(n+K_p-K)!}\times\left(\cos\frac{\beta}{2}\right)^{2J+K-K_p-2n}\left(\sin\frac{\beta}{2}\right)^{2n+K_p-K}. \quad (3\text{-}54)$$

The summation is from the larger of 0 or $(K-K_p)$ to the smaller of $(J-K)$ or $(J-K_p)$. The angle β is the angle between the vector ϱ and z principal axis, i.e.,

**** Hecht and Dennison[20] chose the torsional function in the form of $M_{Kv\sigma}(\alpha)=e^{i(1-\rho)K\alpha}Q_{Kv\sigma}(\alpha)$, which leads to the same integrals $I_{K'v'\sigma}{}^{Kv\sigma}$ in different forms. The form given above is the same as Itoh's.[31]

†††† See p. 38 of reference 20. In these matrices the elements nondiagonal in v have been neglected because they do not contribute to the splittings of the A and E levels but only cause a shift in frequencies.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

the angle between the c' and the z axis:

$$\sin\beta=\rho_y/\rho.$$

For small angles of β Eq. (3-54) can be expanded (to the order of β^2) with the results[62]:

$$(S_1)_{KK}=1-\frac{\beta^2}{4}[J(J+1)-K^2],$$

$$-(S_1)_{K,K+1}=+(S_1)_{K+1,K}=\frac{\beta}{2}[(J-K)(J+K+1)]^{\frac{1}{2}}, \tag{3-55}$$

$$(S_1)_{-1,1}=(S_1)_{1,-1}=\frac{\beta^2}{8}J(J+1),$$

$$(\beta\approx\rho_y/\rho,\ \alpha\approx\rho_z/\rho).$$

4. Energy levels and spectra.—Since the energy matrix given in Eq. (3-49) is diagonal in σ, one may expect to have three series of energy levels corresponding to $\sigma=0$, 1, -1. From Eq. (3-46) the energy submatrices for $\sigma=1$ and $\sigma=-1$ are identical, therefore the energy levels for $\sigma=\pm1$ are doubly degenerate. The eigenfunctions for $\sigma=\pm1$ correspond to the symmetry species E of the C_3 group, and the levels with $\sigma=0$ have the symmetry properties of the A species. Since v is nearly a good quantum number (the elements off-diagonal in v are small), a torsional level is labeled by v and σ, and associated with each torsional level one has a set of rotational energy levels characterized by the quantum numbers J, K, M.

The selection rules $\Delta\sigma=0$ (derived in Sec. 6) predict two lines for each rotational transition associated with the ground torsional state $v=0$ for relatively high barriers, one from the nondegenerate torsional levels $\sigma=0$, and one from $\sigma=\pm1$. Figure 4 shows two pairs of such transitions arising from the two K doublets of a slightly asymmetric top. The splitting of the doublets depends mainly on the tunneling, and the separation between the two pairs of doublets depends on the asymmetry. When the splitting and the separation become comparable in magnitude a mixing of states occurs such that "forbidden lines" can now occur. This is discussed in Sec. 5.

The doublet structure in the microwave spectra is characteristic of the threefold high barrier case, and when the splitting is not too large, the doublets can easily be recognized because the Stark effect is virtually identical for the two lines. Since the δ's and the integrals I are to a good approximation independent of K and σ, the difference between the two corresponding eigenvalues of the degenerate and nondegenerate submatrices for a given J arise only from the terms containing $E_{Kv\sigma}$. Therefore, the splittings of the doublets in the microwave spectrum depend to the first order only on $E_{Kv\sigma}$ but not $\delta_{K'v\sigma}{}^{Kv\sigma}$ and $I_{K'v\sigma}{}^{Kv'\sigma}$ which cause only a shifting in the positions of the lines. The effect of the δ's and I's can be included,[20] but the relations are rather complicated. After the transformation S, the resulting Hamiltonian contains some elements which carry the δ's as coefficients and have the same structure as the ordinary rigid asymmetric rotor matrix elements.[62] As a result the δ's add a correction to the rotational constants which is the same for the degenerate and nondegenerate levels.

Since the energy levels (except $K=0$, $\sigma=0$) are all degenerate for symmetric tops with high barriers, it can be seen that the presence of the asymmetry splits some of these degenerate levels so that the levels with $\sigma=0$ are now nondegenerate. The splitting of the levels by asymmetry can be seen in the following way: to split a pair of degenerate levels of the unperturbed system (the symmetric molecule), say $K=2$, $\sigma=0$, and $K=-2$, $\sigma=0$, it is necessary that the perturbation (the asymmetry) offers matrix elements connecting these two levels directly or through high-order terms. In the present case these two levels do indeed interact with each other through the asymmetry terms as

$$K=2,\ \sigma=0\leftrightarrow K=0,\ \sigma=0\leftrightarrow K=-2,\ \sigma=0.$$

On the other hand, since the asymmetry terms are diagonal in σ, the pairs of levels with quantum number K, $\sigma=1$, and $-K$, $\sigma=-1$ which are degenerate in the symmetric top case, can never connect each other under any high-order interaction. Consequently all the levels with $\sigma=\pm1$ remain doubly degenerate. This is shown in Fig. 4 where the thickness of the levels was drawn to illustrate the degeneracy. Each J, K level has a weight of six from the two K levels and the three internal rotor levels.

At the limit of infinitely high barriers, the δ's and I's approach zero and the energy matrix in Eq. (3-49) reduces to that of a rigid rotor.

5. Splitting of the hindered rotation doublets.—The splittings of doublets associated with the ground torsional level are utilized to determine the barrier height. Hecht and Dennison[20] have given expressions for the differences between the A and E component of a rotational state in the ground torsional level up to $J=3$. With these formulas the barrier can be evaluated from the splittings of the low J rotational lines. For higher J levels the splitting of the A and E levels can be calculated as follows: A transformation S_1S_2' is applied to the Hamiltonian where S_1 corresponds to a rotation of the axes as defined previously and S_2' is the matrix which changes the basis vectors from the symmetric top wave function to the Wang symmetric and antisymmetric combinations. Note that S_2' is different from S_2. Let the transformed Hamiltonian be written as $H_R+H_T+H_{IR}$ where H_R is the rotational part, H_T is the internal rotation part dependent on $E_{Kv\sigma}$, and H_{IR} is the internal rotation part dependent on the integrals which, as pointed out before, have little effect on the splittings of the internal rotation doublets. For slightly asymmetric tops the splittings are caused primarily by

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

the difference in the diagonal elements of H_T between the A and E levels provided the asymmetry splitting is not too small.

Approximate formulas for the diagonal elements in the H_T were derived by Lide and Mann[62] for the case of small values of β. These formulas for the splitting of the energy levels follow directly from Eq. (3-55) and are given in Eq. (3-56):

$$\begin{aligned}\Delta E_{JK}&=\Delta_K+(\beta^2/4)[(J+K)(J-K+1)\Delta_{K-1}\\&\quad-2[J(J+1)-K^2]\Delta_K\\&\quad+(J-K)(J+K+1)\Delta_{K+1}],\\\Delta E_{J1}&=\Delta_1+(\beta^2/4)[(J-1)(J+2)(\Delta_2-\Delta_1)],\\&\quad(J+K_{-1}+K_{+1}\text{ even}),\\\Delta E_{J1}&=\Delta_1+(\beta^2/4)[2J(J+1)\Delta_0\\&\quad-(3J^2+3J-2)\Delta_1+(J-1)(J+2)\Delta_2],\\&\quad(J+K_{-1}+K_{+1}\text{ odd}),\\\Delta E_{J0}&=\Delta_0+(\beta^2/2)J(J+1)(\Delta_1-\Delta_0),\end{aligned}\tag{3-56}$$

where

$$\Delta_K=-\tfrac{1}{2}(E_{Kv1}+E_{Kv-1})+E_{Kv0}\approx+\tfrac{3}{2}Fa_1\cos\left(\frac{2\pi}{3}\rho K\right)$$

$$=\Delta_0\cos\left(\frac{2\pi}{3}\rho K\right),\tag{3-57a}$$

$$\Delta_0=+\tfrac{3}{2}Fa_1=+(9/4)F\Delta b.\ddagger\ddagger\ddagger\ddagger\tag{3-57b}$$

In general the off-diagonal elements of H_T should also be considered for the E levels especially when the asymmetry splitting is small. They connect the states corresponding to the symmetric and antisymmetric combination of the symmetric top functions and have the form of

$$(K_+|H|K_-)=\delta_K=+\frac{\alpha}{2}(E_{Kv\sigma}-E_{-Kv\sigma})$$

$$=+\frac{\alpha}{\sqrt{3}}\Delta_0\sin\left(\frac{2\pi}{3}\rho K\right),\tag{3-58}$$

where

$$\alpha=\rho_z/\rho=(1-\beta^2)^{\frac{1}{2}}.$$

Both the quantities Δ_K and δ_K can be approximated by the trigonometric form only when the barrier is relatively high such that $E_{Kv\sigma}$ can be approximated accurately by the Fourier coefficients a_0 and a_1 with the neglect of the higher terms. The term δ_K is equivalent to the linear terms in P_z in the PAM and the term Δ_K is equivalent to the shift in the rotational constants, i.e., $C_{v0}-C_{v1}$.

The asymmetry term is of the form

$$\delta_{Asy}=-\tfrac{1}{4}(A-B)[(J-K)(J-K-1)\times(J+K+1)(J+K+2)]^{\frac{1}{2}}.\tag{3-59}$$

For the $K=1$ levels only, the asymmetry and the off-diagonal elements of H_T compete directly with one another for the shifting of the levels. If the two effects are of the same order of magnitude, then the correction can be determined from the

$$\tfrac{1}{2}(4\delta_K{}^2+\delta_{Asy}{}'^2)^{\frac{1}{2}}.\tag{3-60}$$

Here $\delta_{Asy}{}'$ is the asymmetry splitting and not the asymmetry term (note that for $K=1$, they are the same). If $\delta_K\gg\delta_{Asy}{}'$, the suitable zeroth-order basis functions are no longer the Wang functions but rather the symmetric rotor functions. After the transformation S_1, the energy matrix can then be diagonalized exactly or by perturbation methods. If, on the other hand, $\delta_{Asy}{}'>\delta_K$, then the splitting due to the off-diagonal elements of H_T must be calculated by the proper asymmetric rotor wave functions, i.e., the effects of β^2 and higher-order terms should be considered. Lide and Mann[62] have derived expressions to the order β^2 for the frequencies of the parallel transition, $J+1, K\leftarrow J, K$.

6. General case.—The most general type of molecules, where the internal rotor is arbitrarily located with respect to the principal axes, has been considered by Pitzer and Gwinn,[84] Hecht and Dennison,[20] and Burkhard.[7] The method follows in an analogous manner, only with added complications. As an aid in the computation it is usually easier to determine the molecular parameters such as ϱ and r by the PAM and then convert them to the IAM.

7. Intermediate barrier case.—If the barrier hindering internal rotation becomes too low for the high barrier approximation to be valid, the matrix elements off-diagonal in v would have appreciable effect on the energy levels. The whole matrix can then be diagonalized either exactly or approximately by using modified perturbation procedure.

Methyl alcohol, methyl amine, and methyl mercaptan are examples of molecules with intermediate barriers, i.e., low s values. Considerable work has been done on these molecules and this is discussed in Sec. 8. The low s values in these molecules are caused in part by the small reduced moment of the methyl group. With the light asymmetric frames, these molecules have only a slight symmetry. All the off-diagonal matrix elements due to asymmetry and internal rotation are small and their effect can be calculated by perturbation theory. For the main Q branch series ($\Delta J=0$, $\Delta K=\pm1$) reported, the transition positions are determined essentially by the limiting symmetric top rotational transition plus the internal rotational transition. The off-diagonal elements contribute small terms which can be expanded in a power series of $J(J+1)$. The coefficients of these power series are the perturbation correction terms including centrifugal distortion. Usually some parallel transitions are needed in addition to aid in the structural determination.

Should the asymmetry be large, the solution of the low reduced barrier problem is quite difficult. With the

‡‡‡‡ Here, Δ_0 is the negative of Δ_0 in reference 20 because we have taken the A level minus the E level. Here Δb is the difference between the values of b [see Eq. (3-4)] for the A and E sublevels of a torsional state.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

matrix elements given in Eq. (3-49) a diagonalization of the low J energy matrices could be made. The $(v|v')$ connections probably could be treated to sufficient accuracy by perturbation calculations. Evaluation of the torsional functions needed for calculating the matrix elements is discussed in Appendix 2.

4. LOW BARRIER APPROXIMATIONS

The barriers of most molecules determined from the microwave spectra are in the category of a "high barrier," i.e., of the order of 0.7–5.0 kcal. Here the "high barrier approximation" (Sec. 3) can be used, in which both a rotational and a torsional equation are solved, and any interaction removed by a transformation. In the low barrier approximation the free internal rotational problem is solved exactly and then the barrier is added as a perturbation. So far two molecules have been treated by this method; they are nitromethane (CH_3NO_2)[106,107] and methyl difluoroboron (CH_3BF_2).[74,75]

The possession of a sixfold symmetry in these two molecules is believed to be the cause for the low value of the barrier. Consider the case of CH_3NO_2. The interaction of a single oxygen with the methyl group gives rise to a barrier to internal rotation in the form

$$\tfrac{1}{2}V_3(1-\cos 3\alpha)+\tfrac{1}{2}V_6(1-\cos 6\alpha)+\tfrac{1}{2}V_9(1-\cos 9\alpha)+\cdots.$$

The second oxygen contributes another potential barrier which is 180° out of phase with the one from the first oxygen. The potential due to both oxygen atoms is then

$$V_6(1-\cos 6\alpha)+V_{12}(1-\cos 12\alpha)+\cdots.$$

The present discussion is therefore confined primarily to molecules with sixfold symmetry. The treatment given can, in principle, be extended to the more general case though the actual analysis might become quite complicated.

I. Free Rotation

A. Symmetric Molecules

For symmetric molecules the difference between the solution by the principal axes method and the internal axis method is not as distinct as in the high barrier problems because the torsional equation with its different boundary conditions is now eliminated.

In the PAM, the Hamiltonian is[121]

$$H=A_x(P_x^2+P_y^2)+C_zP_z^2+Fp^2-2C_zpP_z, \quad (4\text{-}1)$$

where

$$A_x=\hbar^2/2I_x=\hbar^2/2I_y=A, \quad (4\text{-}2a)$$

$$C_z=\hbar^2/[2(I_z-I_\alpha)], \quad (4\text{-}2b)$$

$$F=\hbar^2/2rI_\alpha=\hbar^2I_z/[2I_\alpha(I_z-I_\alpha)]. \quad (4\text{-}2c)$$

This equation can be rearranged to Eq. (4-3),

$$H=A(P_x^2+P_y^2)+CP_z^2+F(p-(I_\alpha/I_z)P_z)^2. \quad (4\text{-}3)$$

But this is just Eq. (2-9) in the IAM.[48] The eigenfunctions of these two equations are

$$S_{JKM}(\theta,\varphi)e^{iK\chi}e^{im\alpha}, \quad (4\text{-}4)$$

where m is an integer. The eigenvalues corresponding to Eq. (4-1) and Eq. (4-3) are then, respectively,

$$E=AJ(J+1)+(C_z-A)K^2+Fm^2-2C_zKm, \quad (4\text{-}5)$$

$$E=AJ(J+1)+(C-A)K^2+F[m-(I_\alpha/I_z)K]^2. \quad (4\text{-}6)$$

However, if the IAM Hamiltonian Eq. (2-20) is used in which χ' and α' are the variables, then the eigenfunctions are

$$S_{JKM}(\theta,\varphi)e^{iK\chi'}e^{im'\alpha'}. \quad (4\text{-}7)$$

The choice of m' is governed by the boundary conditions, i.e., the wave function must be invariant under the transformation

$$\chi_1\rightarrow\chi_1+2\pi n_1 \quad \chi_2\rightarrow\chi_2+2\pi n_2. \quad (4\text{-}8)$$

With Eq. (2-15) and following the same method as used in Eq. (3-20) we obtain the result

$$m'=l-(I_\alpha/I_z)K, \quad (4\text{-}9)$$

where l is an integer. Consequently the eigenvalues given in Eq. (4-10) are obtained:

$$E=AJ(J+1)+(C-A)K^2+F[l-I_\alpha/I_z)K]^2. \quad (4\text{-}10)$$

Clearly, if we identify l with m, Eq. (4-10) is identical to Eq. (4-6).

The selection rules for the microwave spectra can be stated as $\Delta J=\pm1$, $\Delta K=0$, $\Delta m=0$ (see Sec. 6) (or $\Delta l=0$). The frequencies of the rotational transitions are

$$\nu_{J+1\leftarrow J}=2A(J+1), \quad (4\text{-}11)$$

which are exactly the same as those for a rigid symmetrical rotor. Therefore, the effect of completely free rotation in symmetric tops cannot be detected from the microwave spectra.

B. Asymmetric Molecules (CH_3NO_2 *type*)

If the methyl group of CH_3NO_2 were allowed to execute free rotation (zero hindering barrier), the microwave spectrum would be considerably different from a rigid molecule. The Hamiltonian may be split into that of a symmetric internal rotor plus the asymmetry term, i.e.,

$$H=H_0+H_1,$$

where

$$H_0=\tfrac{1}{2}(A+B)(P_x^2+P_y^2)+C_zP_z^2+Fp^2-2C_zpP_z, \quad (4\text{-}12a)$$

$$H_1=\tfrac{1}{2}(A-B)(P_x^2-P_y^2). \quad (4\text{-}12b)$$

By using the basis functions Eq. (4-4) which diagonal-

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

ize H_0, we obtain the matrix elements

$$(JKm|H|JKm)=\tfrac{1}{2}(A+B)J(J+1)+[C_z-\tfrac{1}{2}(A+B)]K^2+Fm^2-2C_zKm,$$

$$(JKm|H|J,K\pm2,m)=-\tfrac{1}{4}(A-B)[J(J+1)-K(K\pm1)]^{\frac{1}{2}}\cdot\times[J(J+1)-(K\pm1)(K\pm2)]^{\frac{1}{2}}. \quad (4\text{-}13)$$

The Hamiltonian is diagonal in J, m, and M and the energy matrix consists of submatrices of the order $2J+1$ for a given set of J, m, and M. As different values of K are connected only in steps of two by the off-diagonal elements, each of the submatrices can be factored into even and odd K blocks. When m is equal to zero, the secular equation degenerates into one identical to that of a rigid asymmetric rotor with the exception that C_z is the rotational constant of the framework alone rather than the entire molecule. For $m\neq0$, the Wang transformation does not lead to further factoring of the secular equation because of the presence of the elements linear in K. As examples, the secular determinant with odd values of K for $J=1$ and $J=2$ are given below. The terms independent of K in the diagonal elements have been omitted since they can be regarded as additive constants for a given J and m.

$$J=1\begin{vmatrix}[C_z-\frac{1}{2}(A+B)]-2mC_z-\lambda & -\frac{1}{2}(A-B)\\ -\frac{1}{2}(A-B) & [C_z-\frac{1}{2}(A+B)]+2mC_z-\lambda\end{vmatrix}=0 \quad (4\text{-}14)$$

$$J=2\begin{vmatrix}[C_z-\frac{1}{2}(A+B)]-2mC_z-\lambda & -\frac{3}{2}(A-B)\\ -\frac{3}{2}(A-B) & [C_z-\frac{1}{2}(A+B)]+2mC_z-\lambda\end{vmatrix}=0 \quad (4\text{-}15)$$

For each m there exists an identical secular equation corresponding to $-m$ and hence the energy levels (with $m\neq0$) are all doubly degenerate. The selection rules (see Sec. 6) for dipole radiation are identical to those of rigid asymmetric tops with the additional restriction of $\Delta m=0$. If the dipole moment is along the z axis, the parity of K cannot change.

From the secular equations (4-14) and (4-15) the frequencies of the $J=2\leftarrow1$ odd K, parallel transitions ($\Delta K=0$) are given by

$$\nu_{2\leftarrow1}{}^{\text{odd }K}=2(A+B)+\lambda_2-\lambda_1=2(A+B)\pm\{[4mC_z{}^2+(9/4)(A-B)^2]^{\frac{1}{2}}-[4m^2C_z{}^2+\tfrac{1}{4}(A-B)^2]^{\frac{1}{2}}\}. \quad (4\text{-}16)$$

For each value of m, two lines are predicted which are symmetrically spaced with respect to the midfrequency of $2(A+B)$. As m increases, the separation between the pair becomes smaller and finally the transitions converge to the band head of $\nu=2(A+B)$. This effect has been observed on CH_3NO_2 by Tannenbaum, Myers, and Gwinn[106,107] and on CH_3BF_2 by Naylor and Wilson.[74,75] The intensities of the higher members would tend to decrease on account of the unfavorable Boltzmann factors.

The other $J=2\leftarrow1$ transition which is even in K, can be treated in a similar manner. As Naylor and Wilson[74,75] have pointed out, the $J=1$, $K=0$ eigenvalue is zero and the $J=2$, $K=0$ eigenvalue is positive, decreasing with increasing m. This group of lines therefore converge toward $2(A+B)$ from the high-frequency side as m increases. For large values of m the energy matrix Eq. (4-13) may be solved by second-order perturbation theory which leads to

$$E=\tfrac{1}{2}(A+B)J(J+1)+[C_z-\tfrac{1}{2}(A+B)]K^2+Fm^2-2C_zKm+\frac{(A-B)^2}{64\{[C_z-\frac{1}{2}(A+B)](K-1)-C_zm\}}\times[J^2-(K-1)^2][(J+1)^2-(K-1)^2]-\frac{(A-B)^2}{64\{[C_z-\frac{1}{2}(A+B)](K+1)+C_zm\}}\times[J^2-(K+1)^2][(J+1)^2-(K+1)^2]+\cdots. \quad (4\text{-}17)$$

The transitions of the type $\Delta J=\pm1$, $\Delta K=0$,§§§§ $\Delta m=0$ again give rise to a band-like structure for each given J.

II. A Low Sixfold Barrier

The sixfold potential barrier introduces to the Hamiltonian an additional term of the form

$$V(\alpha)=(V_6/2)(1-\cos6\alpha)=(V_6/4)(2-e^{i6\alpha}-e^{-i6\alpha}). \quad (4\text{-}18)$$

The first term $\frac{1}{2}V_6$ merely represents a common additive constant and can be safely dropped. The energy matrix analogous to Eq. (4-13) is

$$(JKm|H|JKm)=\tfrac{1}{2}(A+B)J(J+1)+[C_z-\tfrac{1}{2}(A+B)]K^2+Fm^2-2mC_zK$$

$$(JKm|H|J,K\pm2,m)=-\tfrac{1}{4}(A-B)\{[J(J+1)-K(K\pm1)][J(J+1)-(K\pm1)(K\pm2)]\}^{\frac{1}{2}} \quad (4\text{-}19)$$

$$(JKm|H|JK,m\pm6)=-\tfrac{1}{4}V_6.$$

When V_6 is small, the nondiagonality in m may be re-

§§§§ In the present case K is nearly a good quantum number.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

moved by a Van Vleck transformation.[34] To the second-order approximation the matrix elements reduce to those given in Eq. (4-13) with the additional diagonal term

$$\frac{V_6^2}{192[C_zK-F(m+3)]}-\frac{V_6^2}{192[C_zK-F(m-3)]}=\frac{FV_6^2}{32[(C_zK-Fm)^2-9F^2]}. \quad (4\text{-}20)$$

The submatrices for m and $-m$ are still identical and the energy levels remain doubly degenerate. Calculations show that the two correction terms from the Van Vleck transformation in Eq. (4-20) have very little effect on the frequency provided $\frac{1}{4}V_6 \ll F$. The spectrum is therefore almost identical to that for free rotation. The validity of Eq. (4-20) depends, of course, on the relative magnitude of V_6 and F. When the potential barrier becomes larger it is necessary to consider the third- and fourth-order terms from the Van Vleck transformation, but it has been indicated[74] that even when the effect of the higher terms may be quite appreciable as far as the positions of the energy levels are concerned, the effect on the frequency may be negligible.

The foregoing treatment is not directly applicable to the $|m|=3$ levels because a near degeneracy exists and it becomes necessary to consider the $m=-3$ and $m=3$ block as a single secular equation. The matrix elements connecting 3 with 9 and -3 with -9 can, however, still be treated by the Van Vleck procedure. An example is furnished by the block corresponding to $J=1$, K odd:

mK	$(-3,-1)$	$(-3,1)$	$(3,-1)$	$(3,1)$
$(-3,-1)$	a	b	$-\frac{1}{4}V_6$	0
$(-3,+1)$	b	a'	0	$-\frac{1}{4}V_6$
$(3,-1)$	$-\frac{1}{4}V_6$	0	a'	b
$(3,1)$	0	$-\frac{1}{4}V_6$	b	a

(4-21)

where

$$a=\tfrac{1}{2}(A+B)+C_z+9F-6C_z-\frac{V_6^2}{192(6F-C)},$$

$$a'=\tfrac{1}{2}(A+B)+C_z+9F+6C_z+\frac{V_6^2}{192(6F+C)},$$

$$b=-\tfrac{1}{2}(A-B).$$

The symmetry about the antidiagonal allows this matrix to be factored through a transformation which is equivalent to choosing wave functions of the type

$$\frac{1}{\sqrt{2}}[\psi(K, m=3)\pm\psi(-K, m=-3)]. \quad (4\text{-}22)$$

The two resulting matrices are

$$\begin{bmatrix} a' & b+\frac{1}{4}V_6 \\ b+\frac{1}{4}V_6 & a \end{bmatrix}, \quad (4\text{-}23a)$$

and

$$\begin{bmatrix} a' & b-\frac{1}{4}V_6 \\ b-\frac{1}{4}V_6 & a \end{bmatrix}. \quad (4\text{-}23b)$$

The double degeneracy discussed in the previous paragraph and in the free-rotor case is now split by the barrier when $|m|=3$. In fact the energy levels corresponding to m equal to whole multiple of three are all nondegenerate (see Fig. 2). The higher the multiple of three, the higher the order of perturbation required to demonstrate the splitting. The method for calculating these splittings is very similar to that used for the λ doubling in diatomic molecules.

There are now four lines for the $J=2\leftarrow 1$ transition with odd K and $|m|=3$. These four lines appear as two pairs, both symmetrically spaced about the band head $\nu=2(A+B)$ and the splittings of these two pairs of lines are extremely sensitive to the barrier height. In CH_3BF_2 the splitting of one of these pairs, which would be practically coincident at the limit of zero barrier, amounts to about 4000 Mc, for a barrier of 14 cal/mole. These splittings are what make it possible to determine the barrier height with a high degree of accuracy from the microwave spectra.

The selection rules are given approximately by

$$\Delta J=0,\pm 1;\quad \Delta K=0;\quad \Delta m=0.$$

Exact selection rules may be obtained by the group theoretical treatment (Sec. 6) or by an explicit evaluation of the transition moments.

In principle three sets of lines should be observed. The first set, which has the structure of the ordinary rigid asymmetric top spectrum, corresponds to $m=0$. Analysis of these lines yields the two principal moments of inertia perpendicular to the axis connecting the internally rotating groups (the $C-N$ or $C-B$ bond)‖ ‖ ‖ ‖ and also the moment of inertia of the framework about this axis. The second set consists of all the lines with $m\neq 0$ and $m\neq 3$. A band structure is observed if a sufficient number of high m transitions is detected. The $|m|=3$ lines constitute the third set and the splittings of these lines are used to determine the barrier height accurately.

III. A Low Threefold Barrier

The problem of a low threefold potential barrier can be handled in essentially the same fashion. The paper by Wilson, Lin, and Lide[121] gives a treatment for a general N-fold barrier. Here, however, there are no

‖ ‖ ‖ ‖ Nuclear spin symmetry (see Sec. 6) in CH_3NO_2 allows only those energy levels with $K-m$ even so that the second set is also needed to determine the moments of inertia.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

degenerate levels like the $|m|=3$ for the sixfold barrier which are connected by matrix elements of the barrier. Therefore, the barrier height cannot be as accurately determined. Recently $CH_3C\equiv CCH_2Cl$ has been studied[53b] and an estimate of the barrier height has been given.

5. CALCULATIONS OF BARRIER HEIGHTS

After the rather lengthy discussions of the high barrier approximation and the low barrier approximation for the IAM and PAM, the casual reader may be a little perplexed as to when the approximations are valid and which representation to use. Consequently, it seems desirable to give a few suggestions in regard to the analyses of microwave spectra, characteristics of the spectra, and methods of calculation.

I. Analyses of Microwave Spectra

As pointed out previously, the effects of rigid internal rotation are not observed in a symmetric top molecule. As a result, all the applications of internal rotational theory have been done on asymmetric molecules. The methods for determining the barrier heights for the symmetric top molecules from the microwave spectra are discussed in Sec. 7.

For convenience the effect of internal rotation in an asymmetric molecule can be subdivided into the following groups according to the magnitude of the parameter s as defined in Eqs. (3-4):

(*1*) *The pseudo-rigid rotor case* ($s>30$).—With a relatively high barrier and large asymmetry splittings, i.e., the asymmetry splittings are larger than the internal rotation splittings, both members of a transition follow a rigid-rotor spectrum. This is especially true of the low K energy levels with the possible exception of the $K=1$ levels.

The splittings due to the internal torsion may be ascribed to the Δ_K term [Eqs. (3-57)] in the IAM or the ΔW term [cf. Eq. (3-38)] in the PAM. Acetyl fluoride[80] is an example of this class of molecules.

For the ground torsional state, the frequencies of the A transitions are generally somewhat higher than those of the corresponding lines of the rigid molecule while the E transitions usually have lower frequencies. The shift of the A line from the "unperturbed position" is approximately twice that of the E line. This follows from the fact that according to Eqs. (3-57) and (3-28), $\frac{1}{2}(E_{Kv1}+E_{Kv-1})$ is approximately equal to $-\frac{1}{2}E_{Kv0}$ or from Eq. (3-33) and Table II, $w_{01}^{(2)}$ is very close to $-\frac{1}{2}w_{00}^{(2)}$. Many exceptions to this rule can occur, especially for the $K=1$ energy levels, but generally it is a useful first approximation.

(*2*) *A departure of the E levels from a pseudo-rigid rotor pattern* ($15<s<30$).—The δ_K term in the IAM or the $w_{01}^{(1)}$ term in the PAM causes the E energy levels to depart from a rigid-rotor pattern and for a given K the two $\Delta K=0$ transitions tend toward one another so that the spectrum becomes similar to that of a slightly asymmetric top. In other words, the internal rotation cancels out part of the asymmetry. This δ_K term is especially important for the $K=1$ energy levels because both the asymmetry and this "nonrigid-rotor term"

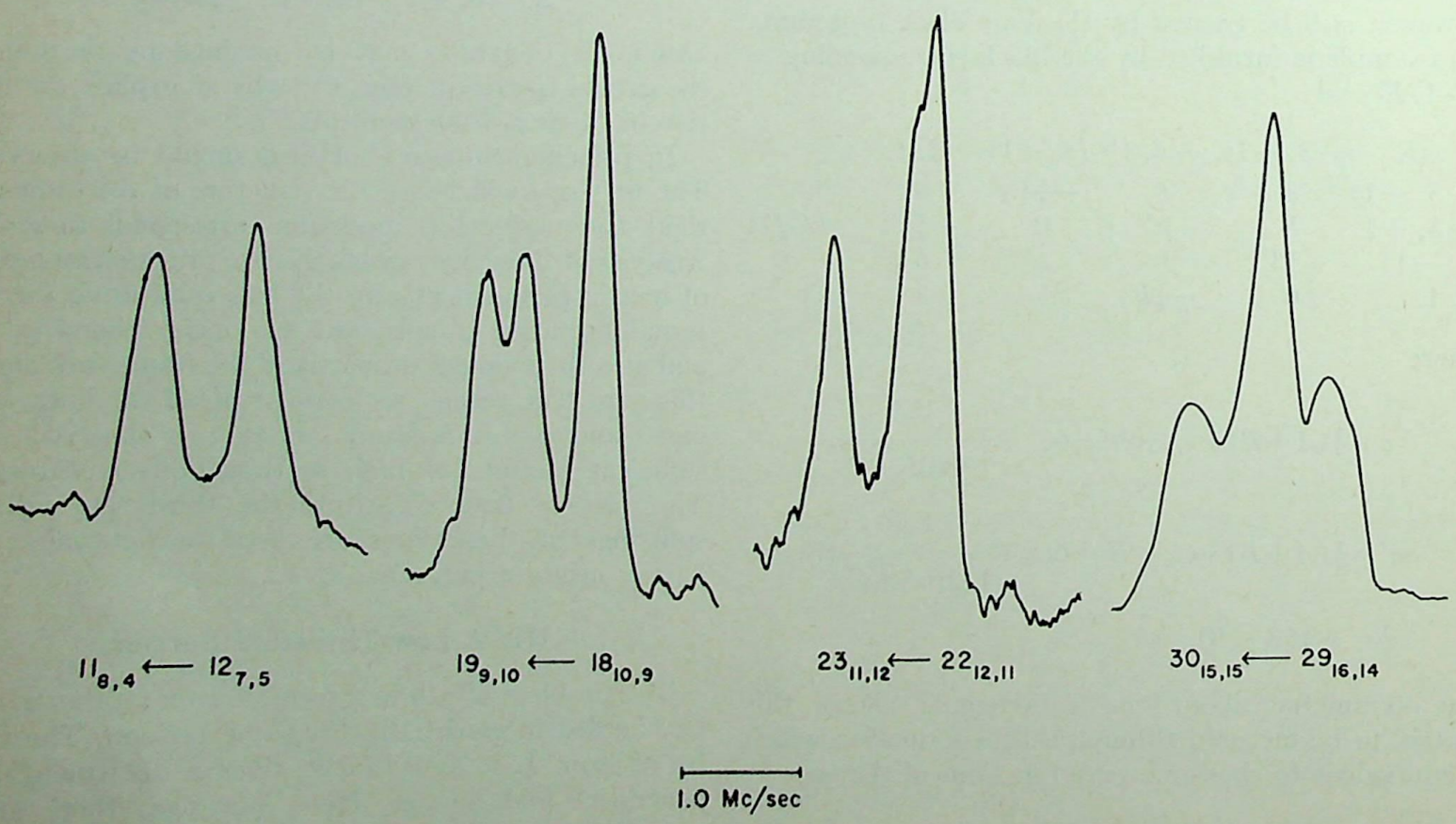


FIG. 5. High K transitions in propylene oxide.[27] The two E lines are superimposed on the left of the first transition but split in the second transition as K increases. In the third transition the E line is now under the A line and finally comes out the other side on the last transition.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

connect the same pair of levels. If either the asymmetry term or the internal rotation term dominates to a large degree, the other can be treated as a perturbation.[62] If the two terms are of the same order of magnitude, then a 2×2 matrix must be solved (cf. Sec. V).

(3) *High K energy levels.*—The high K energy levels of an asymmetric rotor are nearly degenerate even for very large values of asymmetry, so the internal rotational splittings often dominate the asymmetry splittings. These high J, high K energy levels are very similar to symmetric rotor levels except for a shift due to the asymmetry. For the parallel type transitions ($\Delta K=0$) very little splittings are observed, but for the perpendicular transitions ($\Delta K=\pm1$) a triplet is observed consisting of an A symmetry line and two E symmetry lines. This is explained in Sec. 6 II. B. Propylene oxide[27] is an example of this group and some of these triplets are shown in Fig. 5.

(4) *Intermediate barrier heights* ($s<15$).—When the internal rotational splittings are of the order of the spacings between different rotational levels, then the various approximations used in the high barrier case are not generally valid. This is true especially for the perpendicular transitions. Consequently, a more detailed calculation is needed. The IAM is usually recommended for the intermediate barrier problem, especially for the slightly asymmetric molecules, e.g., methyl alcohol. As a guide to the first analysis, the low J parallel transitions should be used to obtain the approximate barrier height and molecular structure since these lines are relatively barrier insensitive. A more detailed analysis of the perpendicular transitions is then made to determine the barrier height more accurately. For molecules with large asymmetry (such as acetic acid[105]) many times the PAM is more convenient.

(5) *Low barrier heights* ($s<2$).—For barrier heights in this region the low barrier approximation given in Sec. 4 is recommended. A considerable number of lines will be observed with all the possible m values. Naturally the Boltzmann factor eventually decreases the intensity of the high m lines.

Generally, the $m=0$ lines should be analyzed first, followed by the low m (except $|m|=3$) transitions which are relatively barrier insensitive. Then an accurate barrier height can be obtained from the $|m|=3$ levels which are widely split by the sixfold potential energy term.

Finally, in the analysis of the spectra, regardless of the s values, the choice of the axis of quantization is gauged by the axis of internal rotation in both the IAM and PAM. This follows directly for the IAM. In the PAM the principal axes closest to the top axes are used as the z axes even if this does not conform to the best choice for the solution of asymmetric rotor energy levels.[41]

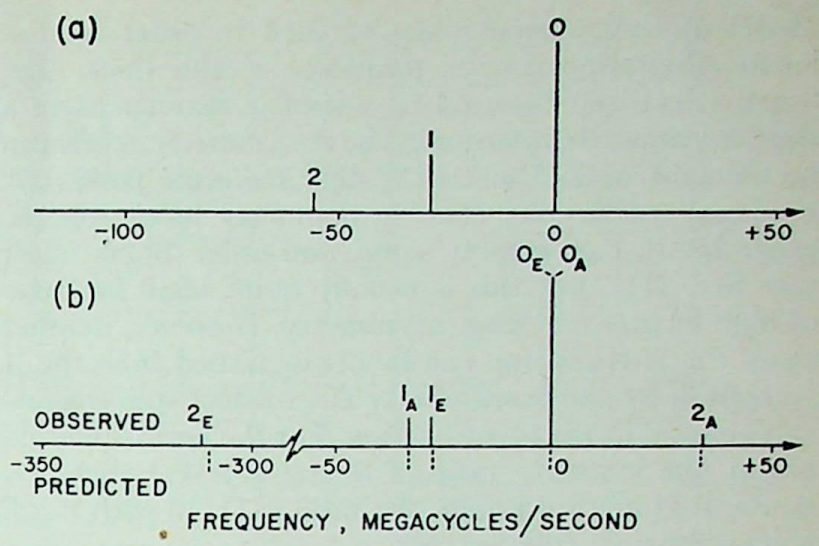


FIG. 6. A rotational line in the ground state and excited vibrational (or torsional) states. (a) The rotational lines of the ground state and first two excited vibrational levels (or torsional levels without internal rotational splitting). (b) The same pattern including internal rotational splitting into the A and E lines. The observed spectrum is from propylene oxide.[27]

II. General Characteristic of the Spectra

While the most intense part of the spectrum for molecules with a relatively high barrier arises from the transitions in the ground torsional state, the rotational transitions in the excited torsional states are usually observable. For example, at room temperature the intensities of the spectral lines from the first excited state could be reduced by a factor of two or three and those from the second excited states by a factor of about ten. These weaker lines (known as satellites) are shifted from the main line associated with the ground torsional state because of the influence of the vibration of the molecule. This total effect is taken into account by using effective rotational constants B_i which are related to the actual rotational constants B_e as $B_i=B_e-\Sigma_i\times(v_i+\frac{1}{2})\alpha_i$. The splittings between the A and E satellite lines in the "heavy" molecules are, however, rather insensitive to the effect of molecular vibrations and, to a first approximation, are caused by the internal rotation only. The general formulas for the A and E splittings developed in Sec. 3 are applicable to the rotational lines in the excited torsional states provided that such torsional levels are far below the top of the potential barrier. As the internal rotation splittings become magnified in the excited torsional states, the potential barrier many times can be determined from the splittings of the satellites.[22] In propylene oxide[27,101] the same barrier can account for the splittings of the rotational lines in the three lowest torsional levels (see Fig. 6). In the lighter molecules, such as methyl alcohol and methyl silane,[42] the splittings of the satellite rotational lines cannot be completely ascribed to the internal rotation. This is because the effects of molecular vibrations contribute appreciably to the splittings as well as the frequency shifts of the rotational lines in the excited states.

In the identification of microwave transitions, a rigid

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

model of the molecule can be used in order to first locate the approximate positions of the lines. The Stark effect can be used in a similar manner as in a rigid asymmetric molecule. The A symmetry levels can be thought of as those of a rigid molecule possessing torsional oscillations. The E symmetry levels, on the other hand, can exhibit some first-order Stark effect (see Sec. III), but this is usually quite small for cases of high barrier and large asymmetry. However, in other cases the E transition can be distinguished from the A transition by the Stark effect. The nuclear spin weights can also aid in the identification. For the normal methyl group, the intensity ratio of A and E is 1:1; but in a deuterated methyl group, the ratio is 11:16 with the E level stronger[38] (see Sec. 6).

With the identification of the various transitions, approximate unperturbed rotational constants can be determined by using either the average position of the A and E doublet or else two-thirds the doublet separation from the A line. The latter method is a better approximation except for the $K=1$ levels. With the approximate rotational constants one can determine the structure of the molecule which along with the A-E splittings may be used to obtain the barrier height. Successive iterations lead to more accurate results.

III. Stark Effect

For the case of free rotation, the lines with $m=0$ show a second-order Stark effect (first order if accidental degeneracy exists) which is a characteristic of a rigid-rotor spectrum. When m is not equal to zero the situation is different. As stated before, the Wang combinations of the K and $-K$ symmetric top functions are no longer the proper zeroth-order wave functions because of the presence of the linear K terms in the secular equation. For example, an eigenfunction for $J=1$ and odd values of K may be written as

$$\Psi=a\psi(K=1)+b\psi(K=-1), \tag{5-1}$$

where a and b are not identical. The expectation values of the direction cosines between the electric field and the body-fixed z axis is then

$$\int\Psi^*\cos\theta\Psi d\tau=\frac{(a^2-b^2)KM}{J(J+1)}. \tag{5-2}$$

This leads us to the conclusion that, in general, a first-order Stark effect is observed provided m, K, and M are all different from zero. For example, with $J=1\leftarrow 0$ only $\Delta K=0$ and $\Delta M=0$ are allowed and the transitions have a second-order Stark lobe while the Stark effect of the $J=2\leftarrow 1$ transitions (with odd K) is first order for $M=\pm 1$ and second order for $M=0$.

When the barrier is low[75,107] the Stark patterns of the lines with $|m|\neq 3$ are practically identical to those for free rotation as the Van Vleck transformation offers only minor modifications to the wave functions. However, for the case of $|m|=3$, the transformation [see Eq. (4-22)] eliminates the diagonal elements of $\cos\theta$. The resulting off-diagonal elements then give rise to a second-order Stark effect with a large Stark coefficient analogous to the case of near degeneracy in rigid asymmetric tops. This serves as a means for identifying the $|m|=3$ lines.

When the barrier becomes higher, the asymmetry has a larger influence. The coefficients a and b in Eq. (5-1) depend on δ_K and δ_{Asy}' (defined in Sec. 3) as

$$a^2-b^2=\pm 2\delta_K(4\delta_K^2+\delta_{\mathrm{Asy}}^2)^{-\frac{1}{2}}. \tag{5-3}$$

Since δ_K is very barrier sensitive, the first-order Stark effect rapidly vanishes as the barrier height is increased.

IV. High Barrier Case

In the symmetric top case, the energy associated with the internal rotation for the IAM is expanded in a Fourier series [Eq. (3-27)] while for the PAM it is expanded in a power series [Eq. (3-13)] as

$$E_{Kv\sigma}=F\left[a_0+a_1\cos\frac{2\pi}{3}(\rho K-\sigma)+a_2\cos\frac{4\pi}{3}(\rho K-\sigma)+\cdots\right], \tag{5-4}$$

$$E_{Kv\sigma}=F[w_{v\sigma}{}^{(0)}+w_{v\sigma}{}^{(1)}\rho K+w_{v\sigma}{}^{(2)}(\rho K)^2+\cdots]. \tag{5-5}$$

In Fig. 7, the various Fourier coefficients are plotted from $s=10$ to $s=100$ on a semilogarithmic scale, and in Fig. 8, the coefficients of the power series in Eq. (5-5) are given. From Eqs. (5-4) and (5-5) the connecting relations between these two types of coefficients can easily be established. They are given in Table III.

TABLE III. The relations between the Fourier coefficients and the power series coefficients [see Eqs. (5-4) and (5-5)].

Factor $N^2/4$	a_0	a_1	a_2	a_3	a_4	a_5
$b(\pi)$	1	1	1	1	1	1
$b(2\pi)$	1	−1	1	−1	1	−1
$b(3\pi)$	1	$-\frac{1}{2}$	$-\frac{1}{2}$	1	$-\frac{1}{2}$	$-\frac{1}{2}$
$b(4\pi)$	1	0	−1	0	1	0
$b(6\pi)$	1	$\frac{1}{2}$	$-\frac{1}{2}$	−1	$-\frac{1}{2}$	$\frac{1}{2}$
$(N^2/4)b(\pi x)$	$a_0+a_1\cos(2\pi/x)+a_2\cos(4\pi/x)+a_3\cos(6\pi/x)+\cdots$					
Threefold case ($N=3$)						
$\frac{3}{2}[b(\pi)-b(3\pi)]$		1	1	0	1	1
$(9/8)[b(\pi)-b(2\pi)]$		1	0	1	0	1
$-2\sqrt{3}p_{vv}/\pi=\sqrt{3}w_{v1}{}^{(1)}/\pi$		1	−2	0	4	−5
$-9w_{v0}{}^{(2)}/2\pi^2$		1	4	9	16	25
$+9w_{v1}{}^{(2)}/\pi^2$		1	4	−18	16	25
$-3\Delta w_v{}^{(2)}/\pi^2$		1	4	0	16	25
$(w_{v1}{}^{(2)}+\frac{1}{2}w_{v0}{}^{(2)})/(3\pi^2)$		0	0	−1	0	0
$-27\sqrt{3}w_{v1}{}^{(3)}/2\pi^3$		1	−8	0	64	−125
$+243w_{v0}{}^{(4)}/2\pi^4$		1	16	81	256	625
$-243w_{v1}{}^{(4)}/\pi^4$		1	16	−162	256	625

e.g., $(N^2/4)b(3\pi)=a_0-\frac{1}{2}a_1-\frac{1}{2}a_2+a_3-\frac{1}{2}a_4-\frac{1}{2}a_5$ and $-[9w_{v0}{}^{(2)}/2\pi^2]=a_1+4a_2+9a_3+16a_4+25a_5+\cdots$.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

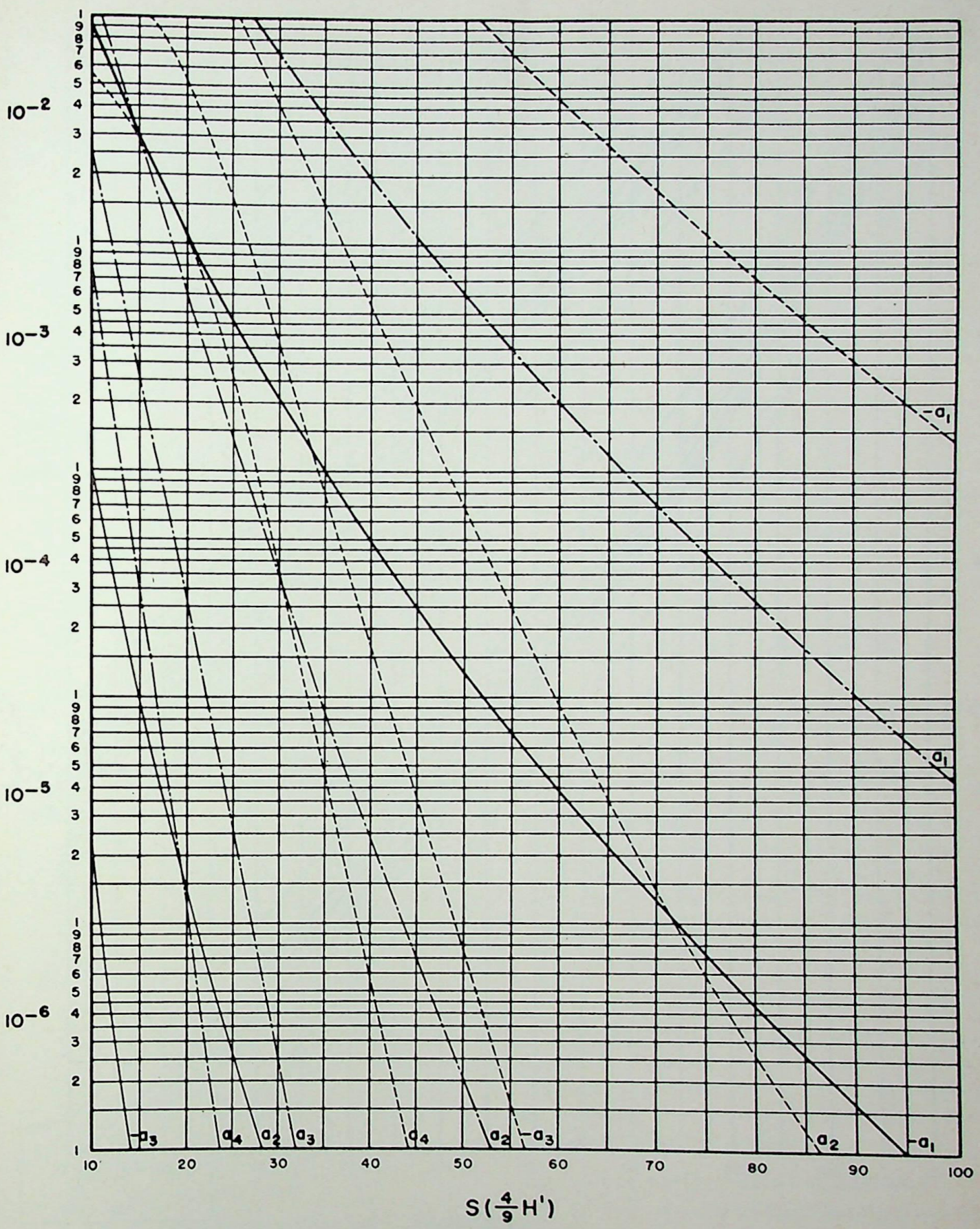


FIG. 7. The Fourier coefficients as a function of the reduced barrier height s for the various torsional levels [see Eq. (5–4)]. ——, $v=0$; —·—, $v=1$; – – – –, $v=2$.

Although these formulas were derived for the symmetric top molecules, they are also applicable to the asymmetric molecules with the necessary modification of F and ρ.

Each of the coefficients (except a_0) are very barrier sensitive, as shown in Figs. 7 and 8. They approximate a functional form of the type $As^{-B}\exp(-C\sqrt{s})$.[20,99] Empirical formulas for the coefficients in Eqs. (5-4) and (5-5) determined by curve fitting are displayed in Table IV. For the high barriers only the Fourier coefficient a_1 is needed from which all the perturbation coefficients of the PAM can be calculated. The domi-

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

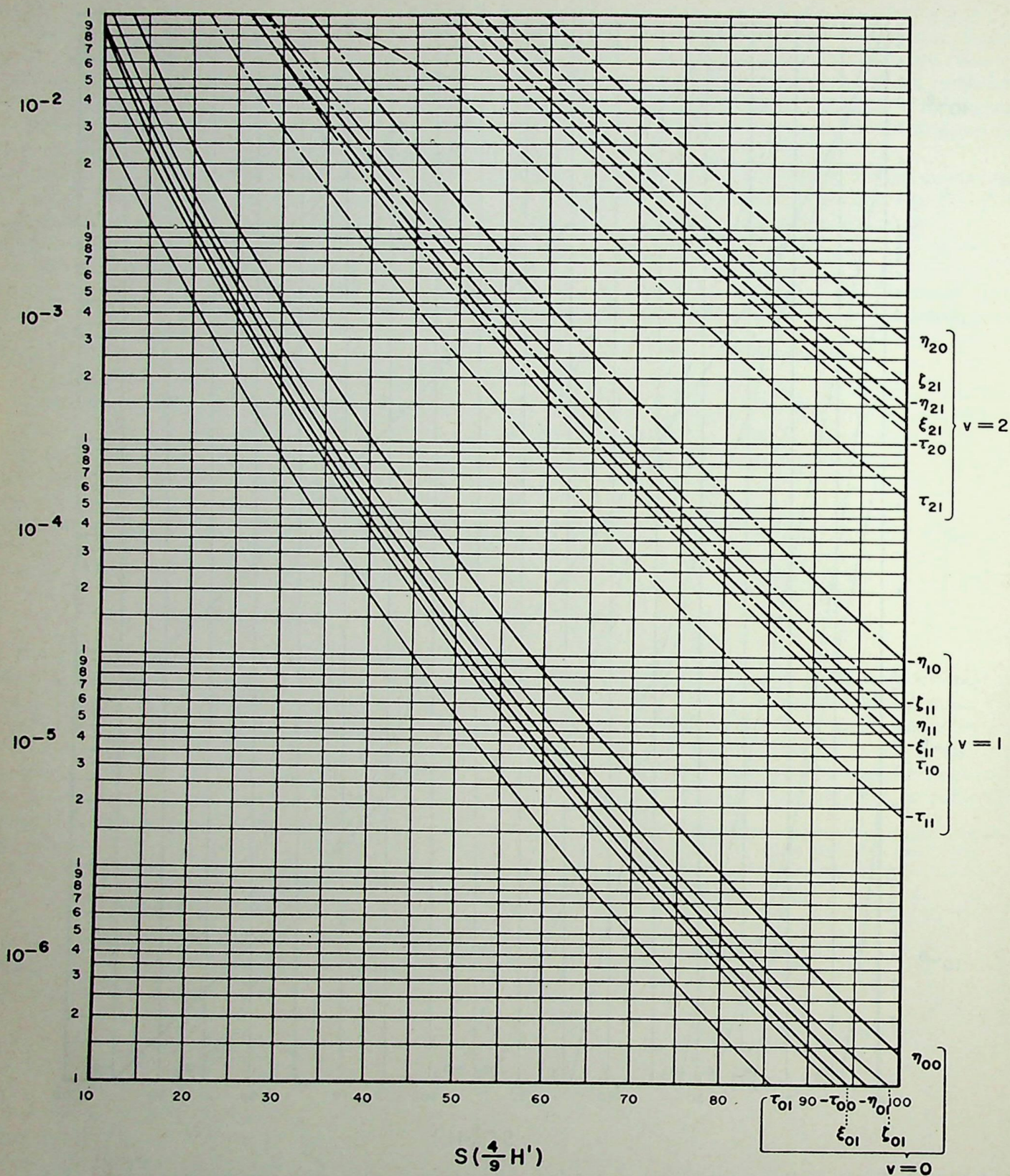


FIG. 8. The power series coefficients as a function of the reduced barrier height s for the various torsional levels [see Eq. (5-5)]. In order to distinguish the various perturbation terms the following notation is used: $w^{(1)}=-2\xi$, $w^{(2)}=\eta$, $w^{(3)}=\zeta$, $w^{(4)}=\tau$. The first subscript denotes the torsional state v, and the second denotes the sublevel σ. ——, $v=0$; —·—, $v=1$; – – – –, $v=2$.

nance of a_1 is seen in Table IV by the fact that the empirical constants B and C have all about the same value for $v=0$. The small deviations are caused by the low s values.

For low values of J, K, and ρ, the agreement between the IAM and PAM is very good. The splittings of the $J=1$ energy levels are given as a comparison in Table V. These results become identical if ρK is small and the

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

barrier is high. Under such circumstances Eq. (5-4) can be approximated by the first two terms and $\sin(2\pi/3)\rho K$ may be replaced by $(2\pi/3)\rho K$. A similar agreement between the IAM and PAM is found for the higher J levels.

These connecting relations also indicate the limitations of the two methods: the IAM is restricted to low J because of computational difficulties, and the PAM is restricted to low values of ρK. If J and ρK are large and the asymmetry is small, a method of solution similar to that used by Burkhard and Dennison[8] on methyl alcohol probably would be the best. If J, ρK, and the asymmetry are all large, both the PAM and IAM involve a considerable amount of calculation. As the barrier is lowered, more expansion terms in Eqs. (5-4) and (5-5) are needed to determine the energy levels.

V. Sample Calculation on the Spectrum of Acetaldehyde (CH_3CHO)

Some of the points previously discussed in regard to the determination of the barrier height from the observed spectrum are illustrated by considering acetaldehyde[38] as an example. The values of the parameters obtained in this section differ slightly from the reported values by Kilb *et al.*,[38] because our calculation is not performed to the same refinement as that of reference 38.

Four of the transitions reported in reference 38 for the normal isotopic species are given in Table VI. The estimated unperturbed frequency given in Table VI was taken as $\frac{2}{3}$ of the doublet separation from the A line. For the $2_{12} \leftarrow 1_{11}$ transition and the $2_{11} \leftarrow 1_{10}$ transition, this procedure is not valid because of the δ_1 term. Here the unperturbed position was estimated from the A line alone. For the $2_{12} \leftarrow 1_{11}$ transition two-thirds of the $1_{01} \leftarrow 0_{00}$ splitting was subtracted from the A line; for the $2_{11} \leftarrow 1_{10}$ transition three times this amount was subtracted from the A line. This procedure follows directly from the PAM (see Table I).

In order to evaluate the barrier height from the internal rotation splittings, the bond lengths and the bond angles of the molecule must be known. These can best be determined from the rotational constants of a number of isotopic species of the molecule.[38] In this sample calculation, however, we use an approximate assumed structure for acetaldehyde. The various molecular parameters calculated from this assumed structure are shown in Table VI.

TABLE IV. Approximate formulas (threefold case).
$\mathrm{Log}X = \log A + B\log s - C(s)^{\frac{1}{2}} + Ds$.

X	v	Range of s	Error	A	B	C	D
$-a_1$	0	20–80	0.1%	0.84469	0.92376	0.89751	8.46×10^{-4}
$-a_1$	1	36–100	0.1%	1.21631	1.89036	0.95594	20.55×10^{-4}
$-a_1$	2	60–100	0.1%	0.66776	3.45716	1.09452	50.19×10^{-4}
$-\Delta b$	0	20–80	0.3%	0.68248	0.85539	0.87703	
$-\frac{1}{2}w_{01}^{(1)}$	0	20–80	0.3%	0.83388	0.83347	0.87430	
$w_{00}^{(2)}$	0	20–80	0.3%	1.18418	0.87348	0.87918	
$-w_{01}^{(2)}$	0	20–80	0.3%	0.89533	0.85530	0.87658	
$\Delta w_0^{(0)}$	0	20–80	0.3%	1.36463	0.86703	0.87825	
$w_{01}^{(3)}$	0	20–80	0.3%	1.02871	0.79882	0.87022	

TABLE V. Comparison of the splitting formulas for an asymmetric molecule with a plane of symmetry.

PAM

$$\Delta E(1_{01}) = \Delta B = F\rho_z^2 \Delta w_v^{(2)}.$$

$$\Delta E(1_{11}) = \Delta C - \frac{(Q_z w_{v1}^{(1)})^2}{A-B} = F\rho_z^2 \Delta w_v^{(2)} - \frac{\delta_1^2}{A-B}.$$

$$\Delta E(1_{10}) = \Delta B + \Delta C + \frac{(Q_z w_{v1}^{(1)})^2}{A-B} = F\rho^2 \Delta w_v^{(2)} + \frac{\delta_1^2}{A-B}.$$

IAM

$$\Delta E(1_{01}) = \Delta_0 + \beta^2(\Delta_1 - \Delta_0) \approx \Delta_0 - \frac{2\pi^2}{9}\rho_z^2 \Delta_0.$$

$$\Delta E(1_{11}) = \Delta_1 - \beta^2(\Delta_1 - \Delta_0) - \frac{\delta_1^2}{A-B} \approx \Delta_0 - \frac{2\pi^2}{9}\rho_z^2\Delta_0 - \frac{\delta_1^2}{A-B}.$$

$$\Delta E(1_{10}) = \Delta_1 + \frac{\delta_1^2}{A-B} \approx \Delta_1 - \frac{2\pi^2}{9}\rho^2\Delta_0 + \frac{\delta_1^2}{A-B}.$$

Connecting relations

$$\beta^2 = \frac{\rho_x^2}{\rho^2}, \quad \alpha^2 = \frac{\rho_z^2}{\rho^2},$$

$$F\Delta w_v^{(0)} = F\Delta b = F(b_A - b_E) = (4/9)\Delta_0 = \tfrac{2}{3}Fa_1,$$

$$-F\Delta w_v^{(2)} = \frac{2\pi^2}{9}\Delta_0 = \frac{\pi^2}{3}Fa_1,$$

$$\Delta_1 - \Delta_0 = \Delta_0\left(\cos\frac{2\pi\rho}{3} - 1\right) \approx -\Delta_0\frac{2\pi^2}{9}\rho^2,$$

$$\delta_1 = +\frac{\alpha\Delta_0}{\sqrt{3}}\sin\frac{2\pi}{3}\rho \approx +\frac{2\pi}{3\sqrt{3}}\rho_z\Delta_0 = \frac{\pi}{\sqrt{3}}\rho_z Fa_1 = F\rho_z w_{v1}^{(1)},$$

$$= Q_z w_{v1}^{(1)} = -2Q_z p_{vv}.$$

A. IAM *Calculations on Acetaldehyde*

From Eqs. (3-56), (3-57), (3-28), and (3-4) we can obtain the expressions for the splitting of the various transitions. For the $J = 1_{01} \leftarrow 0_{00}$ transition a splitting of +3.0 Mc was measured. The approximate formula for this transition is as follows:

$$\nu_A - \nu_E = \Delta\nu = \beta^2(\Delta_1 - \Delta_0) = -0.22134\Delta_0\beta^2$$
$$= -(9/4)F\Delta b\beta^2(0.22134) = -678.5\Delta b \text{ Mc}.$$

By equating this to 3.0 Mc, we obtain a value of -0.004422 for Δb. From the formulas in Table IV, from the curve in Fig. 7 for a_1 ($\Delta b = +2a_1/3$), or from references 3b and 104, the corresponding s value can be determined. A value of 22.943 was obtained which corresponds to a barrier height of 407.7 cm^{-1} (1166 cal/mole). The calculations for the $2_{02} \leftarrow 1_{01}$ follow in a similar manner.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

TABLE VI. Parameters for acetaldehyde.

1. Spectrum

	Obs frequency	Estimated unperturbed frequency
$1_{01}-0_{00}$	19 265.3 (*A*) 19 262.3 (*E*)	19 263.3
$2_{12}-1_{11}$	37 464.6 (*A*) 37 687.2 (*E*)	37 462.6
$2_{02}-1_{01}$	38 512.3 (*A*) 38 506.4 (*E*)	38 508.4
$2_{11}-1_{10}$	39 594.9 (*A*) 39 362.5 (*E*)	39 588.9

2. Rotational constants

C (*A*) 56.9 kMc
B (*B*) 10 163.2 Mc
A (*C*) 9100.0 Mc

The rotational constant notation in parenthesis refers to reference 38.

3. Assumed molecular structure

d_{CH}	1.09 *A* (methyl group)	<HCH	108°15′
d_{CC}	1.54	<CCH	120°
d_{CO}	1.21	<CCO	120°
d_{CH}	1.09 (aldehyde group)		

4. Moments of inertia

I_x	55.553 amu A^2	I_{aa}	55.553
I_y	49.742	I_{bb}	43.176
I_z	8.890	I_{cc}	15.086
I_α	3.079	I_{bc}	14.563

5. Internal rotation parameters

λ_y	0.3910	$F=236.9$ kMc$=17.77$ cm^{-1}
λ_z	0.9204	$\Delta_1=0.77866\Delta_0$
r	0.6931	$\Delta_2=0.21262\Delta_0$
β^2	0.00583	
ρ	0.31941	$(2\pi/3)\rho K=38.86$°K

In the calculations of the barrier height from the $2_{12}\leftarrow 1_{11}$ transition and the $2_{11}\leftarrow 1_{10}$ transition, the δ_1 term must also be considered. For the 2_{12} and 2_{11} energy levels we have the following 2×2 matrix (formed by the K and $-K$ state) with the eigenvalues $\pm\lambda_2$:

$$\begin{bmatrix} 190\,400\Delta b & 1594.8 \\ 1594.8 & -190\,400\Delta b \end{bmatrix},$$

where

$$\delta_1=(1/\sqrt{3})\Delta_0\sin(2\pi/3)\rho K=190\,400\Delta b \text{ Mc}$$

from Eq. (3-58);

$$\delta_{Asy}=-\tfrac{1}{4}(A-B)J(J+1)=\tfrac{1}{4}(1063.2)J(J+1),$$

from Eq. (3-59);

and for the 1_{10} and 1_{11} energy levels we have the following matrix with eigenvalues $\pm\lambda_1$:

$$\begin{bmatrix} 190\,400\Delta b & 531.6 \\ 531.6 & -190\,400\Delta b \end{bmatrix}.$$

Here, terms on the diagonal which are common to both states have been omitted. The expression for the splitting in $2_{12}\leftarrow 1_{11}$ transitions from Eq. (3-56) is then

$$\Delta\nu=\beta^2(\Delta_2+2\Delta_0-3\Delta_1)+\delta_1 \text{ terms}$$
$$=378.01\Delta b+(\lambda_2-1594.8)-(\lambda_1-531.6).$$

By equating this to the observed splitting of -222.6 Mc we obtain a Δb of -0.003993 after successive approximations. This leads to a value of 23.56 for s and 418.7 cm^{-1} (1197 cal/mole) for the barrier height. In a similar manner the expression for the $2_{11}\leftarrow 1_{10}$ transition is

$$\Delta\nu=\beta^2(\Delta_2-\Delta_1)+\delta_1 \text{ terms}$$
$$=-1759.0\Delta b-(\lambda_2-1594.8)+(\lambda_1-531.6).$$

By equating this to the observed splitting of $+232.4$ Mc we obtain a Δb of -0.00405. A value of 23.48 is obtained for s and 417.2 cm^{-1} (1193 cal/mole) for the barrier height. These results are compared in Table VII with the results obtained in reference 38 and by the PAM.

B. PAM Calculations on Acetaldehyde

From Table II and Eq. (3-33) we can obtain the expressions for the splittings of the various transitions. The equation for the $1_{01}\leftarrow 0_{00}$ transition is

$$\Delta\nu=\Delta B=\frac{h}{8\pi^2}\frac{\lambda_y{}^2 I_\alpha}{rI_y{}^2}\Delta w_0{}^{(2)}=138.74\Delta w_0{}^{(2)}.$$

By using the approximate formulas in Table IV or the curves in Fig. 8 we obtain an s value of 22.95 and a barrier height of 407.8 cm^{-1} (1166 cal/mole). The transition $2_{02}\leftarrow 1_{01}$ follows in a similar manner.

In order to obtain the splittings for the $2_{12}\leftarrow 1_{11}$ transition and the $2_{11}\leftarrow 1_{10}$ transition it is necessary to diagonalize two 2×2 matrices because of the linear terms in K. These matrices are

$$\begin{bmatrix} -151\,040p_{00}+7680w_{01}{}^{(3)} & 1594.8 \\ 1594.8 & 151\,040p_{00}-7680w_{01}{}^{(3)} \end{bmatrix},$$

and

$$\begin{bmatrix} -151\,040p_{00}+7680w_{01}{}^{(3)} & 531.6 \\ 531.6 & +151\,040p_{00}-7\,680w_{10}{}^{(3)} \end{bmatrix}.$$

TABLE VII. Barrier height in acetaldehyde by various methods.

	IAM	PAM	KLW (38)
$1_{01}-0_{00}$	407.7 cm^{-1}	407.8	408.9
$2_{02}-1_{01}$	409.7	410.1	[b]
$2_{12}-1_{11}$	418.7	418.0	417.3
$2_{11}+1_{10}$	417.2	418.0	419.0
δ_1	-765 Mc	-763 Mc[a]	-760 Mc[c]
V_3 (av)	414 cm^{-1}		

[a] The third-order correction included.
[b] No individual value was reported for this transition in reference 38.
[c] A sixfold potential energy term was included. See the footnote at the end of Sec. 3 II.A3(b).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

Here the third-order terms have been included. The expressions for the $2_{12} \leftarrow 1_{11}$ transition and for the $2_{11} \leftarrow 1_{10}$ transition from Table II are

$$\Delta\nu = \Delta B + (\lambda_2 - 1594.8) - (\lambda_1 - 531.6)(2_{12} \leftarrow 1_{11}),$$
$$\Delta\nu = 3\Delta B - (\lambda_2 - 1594.8) + (\lambda_1 - 531.6)(2_{11} \leftarrow 1_{10}).$$

By equating these to the observed splittings we obtain the following parameters after successive approximations:

$$p_{00} = 5.46 \times 10^{-3} \qquad Q_z w_{01}^{(1)} = -824.7 \text{ Mc}$$
$$\Delta w_0^{(2)} = 1.92 \times 10^{-2} \qquad \Delta B = 2.66 \text{ Mc}$$
$$w_{01}^{(3)} = 8.10 \times 10^{-3} \qquad F(Q_z/F)^3 w^{(3)} = 62.2 \text{ Mc}.$$

From the curves in Fig. 8 or the approximate formulas in Table IV, a value of 23.52 is obtained for s which corresponds to a barrier height of 418.0 cm^{-1} (1195 cal/mole). The results are summarized in Table VII.

6. SELECTION RULES AND NUCLEAR SPIN STATISTICS

I. Selection Rule for Symmetric Molecules

The selection rules for symmetric molecules can be derived easily from the wave function of the internal rotor. Consider, for instance, the wave function obtained from the IAM

$$S_{JKM}(\theta,\varphi) e^{iK\chi} e^{-ip_K \alpha} P_{K\nu\sigma}(\alpha).$$

Since the orientation of the dipole moment is independent of the angle α, the usual selection rules for the dipole radiation of a rigid symmetric rotor are applicable, i.e.,

$$\Delta J = 0, \pm 1; \quad \Delta K = 0.$$

In addition, it may be shown from Eq. (3-25) that the transition moments are different from zero only when $\Delta\sigma = 0$.

II. Selection Rules for Asymmetric Molecules

The selection rules for an asymmetric molecule in the quantum number J is $\Delta J = 0, \pm 1$. This follows from the fact that the total angular momentum for a rotating molecule remains a good quantum number even when internal rotation is present. The selection rules governing the changes of the other quantum numbers are more complicated. In general, these rules can be derived by group-theoretical methods.

The general selection rules are derived from symmetry considerations. Application of these rules to the cases of high barrier and low barrier are then discussed in detail.

A. General Selection Rules

The Hamiltonian of a rigid asymmetric rotor is invariant under the operations

$$\begin{aligned} E&: P_x \to P_x,\ P_y \to P_y,\ P_z \to P_z, \\ C_z&: P_x \to -P_x,\ P_y \to -P_y,\ P_z \to P_z, \\ C_x&: P_x \to P_x,\ P_y \to -P_y,\ P_z \to -P_z, \\ C_y&: P_x \to -P_x,\ P_y \to P_y,\ P_z \to -P_z, \end{aligned} \tag{6-1}$$

and therefore has the symmetry of the four group¶¶¶¶ V.

When the effect of internal rotation is included, the Hamiltonian is no longer invariant under the operations given in (6-1) because of the cross terms between the P's and p [e.g., Eq. (2-25) or Eq. (2-37)] or between the P's and functions of the angle α [e.g., Eq. (2-29) or Eq. (2-31)]. The group properties of the various types of molecules are discussed in the following sections.

1. Molecules with a planar frame.—The Hamiltonian Eq. (2-31) or Eq. (2-37) (with $Q_x = 0$) is invariant under the following operations:

$$\begin{aligned} E&: P_{x,y,z} \to P_{x,y,z},\ \alpha \to \alpha(p \to p), \\ C_x&: P_x \to P_x,\ P_{y,z} \to -P_{y,z},\ \alpha \to -\alpha(p \to -p), \\ C_3&: P_{x,y,z} \to P_{x,y,z},\ \alpha \to \alpha + (2\pi/3)(p \to p), \\ C_3^2&: P_{x,y,z} \to P_{x,y,z}, \\ &\qquad \alpha \to \alpha + (4\pi/3) = \alpha - (2\pi/3)(p \to p). \end{aligned} \tag{6-2}$$

(The operations $C_x C_3$ and $C_x C_3^2$ follow directly from these operations.) These symmetry operations satisfy the properties of the group D_3, and therefore the wave functions have the symmetry of the species A_1, A_2, or E.

For this type of molecule the dipole moment lies on the plane of the framework and has the transformation properties of A_2. The selection rules for the dipole transitions are then[20]

$$A_1 \leftrightarrow A_2, \quad E \leftrightarrow E.$$

2. Molecules with no symmetry in their framework.—For this class of molecules only the operations E, C_3, C_3^2 leave the Hamiltonian unaltered. The energy states then fall into the species A or E of the C_3 group. Since the dipole moment has the symmetry property of A, the selection rules can be summarized as $A \leftrightarrow A$, $E \leftrightarrow E$.

3. Molecules with a frame of C_{2v} symmetry (CH_3NO_2 *type*).[121]—The group properties are of particular interest because the cross term between **P** and p appears only as pP_z. The Hamiltonian is now invariant under the following operations:

$$\begin{aligned} E&: P_{x,y,z} \to P_{x,y,z},\ \alpha \to \alpha(p \to p), \\ C_x&: P_x \to P_x,\ P_{y,z} \to -P_{y,z},\ \alpha \to -\alpha(p \to -p), \\ C_y&: P_y \to P_y,\ P_{x,z} \to -P_{x,z},\ \alpha \to -\alpha(p \to -p), \\ C_z&: P_z \to P_z,\ P_{x,y} \to -P_{x,y},\ \alpha \to \alpha(p \to p). \end{aligned} \tag{6-3}$$

Because of the sixfold symmetry, the Hamiltonian is also invariant under the operation $\alpha \to \alpha + (2n\pi/6)$.

¶¶¶¶ The operation $P_x \to -P_x$, $P_y \to P_y$, $P_z \to P_z$, does not alter the Hamiltonian but does change the commutation relations between the three components of the angular momentum, so it is not considered to be a symmetry operation.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

TABLE VIII. Character table for the point group D_{6h}.

	E	$2C_3$	$3C_2C_3$	C_2	$2C_2C_3$	$3C_yC_3$	C_2	$2C_6$	$3C_2C_6$	C_2C_2	$2C_2C_6$	$3C_yC_6$
A_e	1	1	1	1	1	1	1	1	1	1	1	1
B_{ze}	1	1	−1	1	1	−1	1	1	−1	1	1	−1
E_{1e}	2	−1	0	2	−1	0	2	−1	0	2	−1	0
B_{xe}	1	1	1	−1	−1	−1	1	1	1	−1	−1	−1
B_{ye}	1	1	−1	−1	−1	1	1	1	−1	−1	−1	1
E_{2e}	2	−1	0	−2	1	0	2	−1	0	−2	1	0
A_0	1	1	1	1	1	1	−1	−1	−1	−1	−1	−1
B_{z0}	1	1	−1	1	1	−1	−1	−1	1	−1	−1	1
E_{10}	2	−1	0	2	−1	0	−2	1	0	−2	1	0
B_{x0}	1	1	1	−1	−1	−1	−1	−1	−1	1	1	1
B_{y0}	1	1	−1	−1	−1	1	−1	−1	1	1	1	−1
E_{20}	2	−1	0	−2	1	0	−2	1	0	2	−1	0

These six operations are denoted by E, C_6, $C_6^2(C_3)$, $C_6^3(C_2)$, $C_6^4(C_3^2)$, and C_6^5, where the operation C_6 represents $\alpha \rightarrow \alpha + (\pi/3)$. These six operations along with the "four-group" operations form a new group of twenty-four elements which is isomorphous to the point group D_{6h}. The character table of this group is given in Table VIII.

The dipole moment for this model would ordinarily be along the z axis. If this is the case, the transition moment will have symmetry B_{ze}. The selection rules for dipole absorption will then be $A_e \leftrightarrow B_{ze}$, $B_{xe} \leftrightarrow B_{ye}$, $E_{1e} \leftrightarrow E_{1e}$, $E_{2e} \leftrightarrow E_{2e}$, $A_0 \leftrightarrow B_{z0}$, $B_{x0} \leftrightarrow B_{y0}$, $E_{10} \leftrightarrow E_{10}$, $E_{20} \leftrightarrow E_{20}$.

B. Selection Rules for High Barrier

Molecules with a planar frame may be considered as an example. The limiting rigid-rotor Hamiltonian belongs to the four-group. As the operation C_x is common to both the D_3 group (of the torsional rotor) and the four-group, the A_1 species of the D_3 group corresponds to A and B_x in the four-group while the A_2 levels of the hindered rotor reduces to the B_y or B_z levels of the limiting rigid molecule. As shown before, the selection rules for the rotational states associated with the nondegenerate torsional levels are $A_1 \leftrightarrow A_2$. In terms of the transitions for the limiting rigid rotor these correspond to A, $B_x \leftrightarrow B_y$, B_z, which are indeed the selection rules for a rigid asymmetric top molecule with dipole moment on the yz plane. Hence, for the internal rotor with high barriers the selection rules governing the transitions between the rotational levels belonging to the nondegenerate torsional levels (the pseudo-rigid rotor lines) may simply be taken as those for the corresponding rigid molecules.

In the case of degenerate torsional levels, according to the general symmetry selection rule $E \leftrightarrow E$, transitions between all the E levels are, in general, possible. This can be understood by considering the effective rotational Hamiltonian for a given torsional level. For the E levels the nonrigid-rotor terms, δ_k, tend to eliminate the asymmetry and the transitions approach those of a symmetric rotor. The eigenfunctions no longer have the symmetry of the four-group, and the symmetry selection rules for the wave functions associated with these symmetry species are not applicable. For the case of high barriers and low K, it may be expected that the transitions in the hindered rotor which are forbidden for the limiting rigid rotor are usually weak. However, if transitions involving higher K values are observed, the effect of the internal rotation can now exceed the asymmetry splitting even in the case of relatively high barriers, and the spectrum becomes very interesting. Consider the perpendicular type transition ($\Delta K = \pm 1$) between two groups of four energy levels corresponding to the A and E levels of a K doublet. At low values of K four lines are expected (two doublets) similar to those shown in Fig. 4. As K increases, all four E transitions are possible and six lines appear (two A and four E). At high values of K the wave function may be approximated by the symmetric top wave function. When this is the case the two A levels of a given K become nearly degenerate and the two A lines merge into a single line. The two E levels, however, remain separated because of the linear term P_z in the effective Hamiltonian Eq. (3-33), and two of the four E lines now become forbidden so that for large values of K one observes a group of three spectral lines—(one A and two E). Such a transition from four lines (low K) to three lines (for high K) has indeed been found by Herschbach and Swalen.[27] In the transition region between these two patterns all four E transitions are observed. Those transitions forbidden in the limit of a rigid rotor can be seen to increase in intensity at the expense of the rigid-rotor lines until finally the triplet pattern results (see Fig. 5).

C. Selection Rules for Low Barrier

So far the only molecules which have been found by the methods of microwave spectroscopy to have low potential barriers to internal rotation are CH_3NO_2 and CH_3BF_2*****; our present discussion is confined mainly to this class of molecules. The symmetry selection rules have been given in Sec. A3. In order to apply these rules to the analysis of the microwave spectrum it is necessary to establish some correlation between the symmetry properties and the rotational quantum numbers.

1. Limiting free rotator.—The Hamiltonian and the energy levels were discussed in Sec. 4. Since the quantum number m is diagonal, the wave function can be expressed as a product of $e^{im\alpha}$ and a function of θ, Φ, and χ. This function may be determined from the transformation matrix which diagonalizes the energy matrix in Eq. (4-13). An inspection of the structure of this matrix reveals the fact that a typical wave function has one of the following forms:

$$\Psi_m^{\text{even } K} = e^{im\alpha} \sum_{\text{even } K} a_K S_{JKM}(\theta,\varphi) e^{iK\chi}, \qquad (6\text{-}4)$$

$$\Psi_m^{\text{odd } K} = e^{im\alpha} \sum_{\text{odd } K} a_K S_{JKM}(\theta,\varphi) e^{iK\chi}, \qquad (6\text{-}5)$$

***** Recently $CH_3C{\equiv}CCH_2Cl$[53a] has been studied and reported to have a low barrier.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

where $S_{JKM}(\theta,\varphi)e^{iK\chi}$ denotes the symmetric top wave function and the summation in K in Eqs. (6-4) and (6-5) extends only from $-J$ to J. As explained in Sec. 4, each energy level with $m\neq 0$ is doubly degenerate. This double degeneracy occurs because the energy matrix Eq. (4-13) remains unchanged when m and K are replaced by $-m$ and $-K$. Now those energy levels where m is a multiple of three (and not zero) are associated with the pairs of symmetry species: A_e, B_{ze}; A_0, B_{z0}; B_{xe}, B_{ye}; or B_{x0}, B_{y0}. Here each pair constitutes the double degeneracy which would be split by a barrier. Those energy levels where m is not a multiple of three are associated with the symmetry species: E_{1e}, E_{2e}, E_{10}, E_{20}. Here the degeneracy is not split by the barrier.

Consideration of the C_z operation ($\chi\rightarrow\chi+\pi$) shows that the wave function Eq. (6-4) remains unchanged, but the wave function Eq. (6-5) changes sign. For the C_6^3 operation ($\alpha\rightarrow\alpha+\pi$) both Eq. (6-4) and Eq. (6-5) remain unchanged if m is even but change sign if m is odd. With the aid of the character Table VI, the symmetry properties of the energy levels characterized by different m and K may be summarized as follows:

	$m(m\neq 0)$	K
A_e, B_{ze}, E_{1e}	even	even
B_{xe}, B_{ye}, E_{2e}	even	odd
A_0, B_{z0}, E_{10}	odd	even
B_{x0}, B_{y0}, E_{20}	odd	odd

The selection rules for m and K are such that the parity is not changed. For the quantum number m the selection rule is even more restricted in the case of free rotation. Since the dipole moment is independent of α and neither wave function Eq. (6-4) nor Eq. (6-5) involve states of different m, we have the selection rule $\Delta m=0$. For the K quantum number, the selection rule is $\Delta K=0$, ± 2, ± 4, etc. Here K is nearly a good quantum number and the transitions with $\Delta K=0$ will in general be stronger than those with $\Delta K=\pm 2, \pm 4$, etc. These latter types of transitions result from a mixing of the K levels in Eqs. (6-4) and (6-5). Note, however, that the Wang functions are no longer the appropriate zero-order wave functions because of the linear term in K on the diagonal of the energy matrix.

The discussion presented above pertains to the energy levels with $m\neq 0$ only. When $m=0$, the exponential factor $e^{im\alpha}$ in Eqs. (6-4) and (6-5) and also the linear term in K in the energy matrix disappear. The secular equation degenerates into that of a rigid rotor. Since the expressions in Eqs. (6-4) and (6-5) with $m=0$ are invariant under the operation C_6^3 and C_3, the energy levels with $m=0$ have the symmetry properties of A_e, B_{xe}, B_{ye}, or B_{ze}. The selection rules for the transitions with this series of energy levels are identical to those for a rigid asymmetric top, i.e., $A_e\leftrightarrow B_{ze}$ and $B_{xe}\leftrightarrow B_{ye}$.

2. Low barrier.—As explained in Sec. 4, the addition of a small barrier term $\frac{1}{2}V_0(1-\cos 6\alpha)$ to the Hamiltonian of a free internal rotor does not greatly alter the spacing of the rotational energy levels except for the ones with m equal to whole multiples of three. The nondiagonal elements in m may be removed by a Van Vleck transformation so that the secular equation can be factored approximately into blocks corresponding to different values of m. The basis functions for this effective energy matrix (for $m\neq 3n$) are then

$$(e^{im\alpha}+ae^{i(m+6)\alpha}+be^{i(m-6)\alpha}+\cdots)S_{JKM}(\theta,\varphi)e^{iK\chi}. \quad (6\text{-}6)$$

The coefficients a and b do not depend critically on K. The final composite wave functions are taken as linear combinations of such basis functions. The structure of the secular Eq. (4-19) again is such that the basis functions of odd and even K do not combine in forming the final wave function. Furthermore, since the coefficients a, b are usually very small, one may still use m as the labeling index for the energy states. The selection rule for m is Δm equals a whole multiple of 6. Since the $\Delta m\neq 0$ transitions are weak and usually beyond the microwave region, the selection rules for the dipole transitions may be given approximately as $\Delta m=0$, $\Delta K=0$, $\Delta J=0, \pm 1$. Here again it is assumed that K is nearly a good quantum number. Remembering that each energy level characterized by (m,K) is degenerate with the one by $(-m,-K)$, one may easily show that the selection rules $\Delta m=6n$ (n being an integer) are equivalent to those derived from the symmetry consideration.

For $m=3$ (or whole multiple of three) the solution of the secular equation becomes more complicated as discussed in Sec. 4 II. As an example the block in the secular equation corresponding to $J=1$, $m=3$, odd K was given in Eq. (4-21). This secular equation can be factored by choosing as the basis functions the type of linear combination given in Eq. (4-22) i.e.,

$$\psi(K,m=3)\pm\psi(-K,m=-3). \quad (6\text{-}7)$$

To the zeroth-order approximation these functions may be expressed as

$$e^{3i\alpha}S_{JKM}(\theta,\varphi)e^{iK\chi}\pm e^{-3i\alpha}S_{J-KM}(\theta,\varphi)e^{-iK\chi}. \quad (6\text{-}8)$$

These two linear combinations have the symmetry property of the A_0 and B_{z0} when K is even; and B_{x0} and B_{y0} for odd K. In the case of the example in Eq. (4-21) the energy eigenfunctions arising from the secular Eq. (4-23a) have the symmetry of B_{z0} while those corresponding to the determinant Eq. (4-23b) belong to A_0 species. The approximate wave functions for these four states are

$$\begin{aligned}
\Psi_{++}&=a[\psi(1,3)+\psi(-1,-3)]\\
&\quad+b[\psi(-1,3)+\psi(1,-3)],\\
\Psi_{+-}&=-b[\psi(1,3)+\psi(-1,-3)]\\
&\quad+a[\psi(-1,3)+\psi(1,-3)],\\
\Psi_{-+}&=c[\psi(1,3)-\psi(-1,-3)]\\
&\quad+d[\psi(-1,3)-\psi(1,-3)],\\
\Psi_{--}&=-d[\psi(1,3)-\psi(-1,-3)]\\
&\quad+c[\psi(-1,3)-\psi(1,-3)].
\end{aligned} \quad (6\text{-}9)$$

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

Here the first and second number inside the parentheses of the functions refer to the values of K and m, respectively. For CH_3NO_2 the difference between the two diagonal elements in Eq. (4-23a) and Eq. (4-23b) are roughly ten times greater than the off-diagonal terms; so the ratios of the coefficients a, b, c, and d in Eq. (6-9) may be estimated by perturbation methods as $|a/b| \approx |c/d| \approx 10$. One may construct four approximate wave functions similar to Eq. (6-9) for the states with $J=2, m=3$, and odd K. The coefficients a, b, c, etc., here are somewhat different from those in Eq. (6-9), but the ratios of these coefficients are still of the same order of magnitude as those for $J=1$. In accordance with the symmetry selection rules $A \leftrightarrow B_z$, $B_x \leftrightarrow B_y$, $o \leftrightarrow o$, $e \leftrightarrow e$, derived previously, there are eight possible transitions for $J=2 \leftarrow 1, m=3$ with odd K. Among these eight lines, four are much more intense than the others, and indeed only the four strong lines are observed in the microwave spectrum. These transitions are

$$\Psi_{++} \leftrightarrow \Psi_{++}, \quad \Psi_{+-} \leftrightarrow \Psi_{+-}, \quad \Psi_{-+} \leftrightarrow \Psi_{-+}, \quad \Psi_{--} \leftrightarrow \Psi_{--}.$$

Similar results are also found in the microwave spectrum of CH_3BF_2.

The selection rules for the transitions involving the energy levels with $m=0$, are $A_e \leftrightarrow B_{ze}$, and $B_{xe} \leftrightarrow_{ye}$.

III. Nuclear Spin Weight

To obtain the relative intensities of the components of the internal rotation multiplets (such as the doublet for the high barrier case), the correct combinations of nuclear spin functions ψ_s with the internal torsional function ψ_t must be found. The method for finding such appropriate nuclear spin functions has been given by Wilson.[118] This procedure will be applied to the various types of nonsymmetric molecules.

A. Molecules with No Symmetry

By following Wilson,[118] we observe that for the protons in the methyl radical, the group of permutation operations which are equivalent to the internal rotations of the molecules is isomorphous with the group C_3, as is the group derived from the internal rotation operations. For CH_3 there are eight nuclear spin functions which can be schematically represented by specifying the quantum number m_s for each nucleus as follows:

$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	(one)
$\frac{1}{2}$	$\frac{1}{2}$	$-\frac{1}{2}$ etc.	(three)
$\frac{1}{2}$	$-\frac{1}{2}$	$-\frac{1}{2}$ etc.	(three)
$-\frac{1}{2}$	$-\frac{1}{2}$	$-\frac{1}{2}$	(one)

These eight spin functions form a reducible representation of the C_3 group. The characters of this representation are $\chi(E)=8$, $\chi(C_3)=2$, $\chi(C_3^2)=2$. This reducible representation then belongs to the species $4A+2E_1+2E_2$.[38] The over-all wave function must be antisymmetric to the interchange of protons, i.e., of species A with respect to the operations of C_3. Since the nondegenerate torsional function is of species A, the symmetry of the composite wave functions is $A \cdot (4A+2E_1+2E_2) = 4A+2E_1+2E_2$; therefore, the statistical weight is four. For the degenerate torsional sublevel the reducible representation formed by the total wave function is $(E_1+E_2) \cdot (4A+2E_1+2E_2) = 4A+6E_1+6E_2$. Consequently the members of a torsional doublet (in the case of high barriers) of a particular rotational transition should have equal intensity.

The twenty-seven nuclear spin states for a CD_3 group form a representation of species $11A+8E_1+8E_2$.[38] Thus the nondegenerate levels yield $A \cdot (11A+8E_1+8E_2) = 11A+8E_1+8E_2$, whereas the degenerate levels give rise to $(E_1+E_2) \cdot (11A+8E_1+8E_2) = 16A+19E_1+19E_2$. The relative intensity of the two members of a doublet is therefore 11/16.

This discussion is also applicable to the molecules with a planar frame such as acetaldehyde.

B. CH_3NO_2 Type Molecules

The method outlined in the previous section may also be applied to find the statistical weights of the degenerate and the nondegenerate levels for this type of molecule. When the barrier is low the separation between the nondegenerate level and a degenerate level—say between a level with $m=0$ and one with $m=1$—is of the order $2Fm$ which may be comparable to kT. Hence the relative intensity of the various members of the multiplet arising from the internal rotation is not simply equal to the ratio of the statistical weights. Rather, the Boltzmann factor must be taken into consideration.

In addition, for the case of CH_2NO_2[16] the wave function must be even with respect to the interchange of the two oxygen ($I=0$). This operation is $C_zC_6^3$. Thus, only the levels with symmetry properties of A_e, B_{ze}, E_{1e}, B_{x0}, B_{y0}, E_{20}[107] can exist in CH_3NO_2[16]. For CH_3BF_2 these species will have one-third the statistical weight of the others because of the fluorine nuclear spin ($I=\frac{1}{2}$) and the requirement of antisymmetry for exchange of the two fluorine nuclei.[121 75].

7. VIBRATION-TORSION-ROTATION INTERACTION

The methods for the determination of the potential barrier described in Secs. 3 and 5 utilize the splittings between the same rotational transitions corresponding to the two ground torsional levels. These methods, however, fail when the splittings of the rotational lines vanish as in the case of symmetric tops (see Fig. 4) or fall below the limit of resolution of the spectrograph for very high barriers. The latter difficulty may be overcome, at least partially, by measuring the splittings of the rotational lines in the excited torsional levels (satellites) which are considerably larger than those in the ground torsional state. Because of the influence of the molecular vibrations on the internal torsion and the

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

over-all rotation, these satellites are shifted from the "main lines" (the lines associated with the ground torsional state). Furthermore, the splittings of the satellites sometimes contain sizeable contributions from the effect of molecular vibration (cf. Sec. 5 II) as well as the internal rotation. With a careful analysis of the interaction of vibration, torsion, and over-all rotation, it is possible to determine the barrier height either from the splittings of the satellites or from the shifts of the satellites from the main line. The latter procedure is particularly significant because it is applicable to symmetric top molecules.

Besides the frequency shifts, the intensities of the satellites are decreased by the Boltzmann factor. From the intensity ratios between the satellites and the main line, one can calculate the energy differences between the ground and excited torsional states from which the barrier height may be determined.

In the present section two methods—the satellite frequency pattern and the relative intensities—for determining potential barriers are presented. Neither of these methods gives as accurate results as the "splitting method" described in the previous sections. Therefore, these methods are used only when the splitting method fails to yield the barrier height or when an analysis of the torsional satellites is desired to study torsion-rotation interaction.

I. Satellite Frequency Pattern

In the discussion of the theory of internal rotation in Sec. 2 the effects of molecular vibration have not been considered. This is justified if the torsional frequency is well below the frequencies of the other vibrations. Since the moments of inertia do not depend directly on the torsional normal coordinate, the only way the torsional mode can affect the moments of inertia is by means of higher-order interactions via the other vibrations of the molecule.[42] When the nonrigidity of the two halves of the molecule is taken into account, some interesting results are obtained. Consider first the symmetric molecules such as methyl silane, CH_3SiH_3. Classically, as the CH_3 group oscillates about its equilibrium position, the centrifugal force produces a slight change in the molecular dimensions. Because both the angular velocity and the potential energy are functions of the internal angles α, the instantaneous configuration of the molecule depends also on α. The rotational constant B determined from the microwave spectra is actually related to the average value of the reciprocal of the moment of inertia for a given torsional and vibrational state. One would expect that the rotational constants and thus the rotational lines associated with the different excited torsional states would be shifted by different amounts (even for the case of a symmetric top). From the pattern of these lines one can then determine the height of the potential barrier. Furthermore, this type of internal torsion-vibration interaction also gives rise to an additional splitting of the pairs of lines associated with the A and E sublevels of a *given* torsional level; but as shown by Kivelson,[42] such a splitting for the ground torsional state of CH_3SiH_3 is less than 0.05 Mc which is experimentally unobservable.

The quantum-mechanical treatment can be carried out in this way: The Hamiltonian, including the rotation, internal torsion, and vibration,††††† according to Kivelson,[42,45b] is written as

$$H=\tfrac{1}{2}\sum_{ij}\mu_{ij}P_iP_j+H_T+H_v, \qquad (7\text{-}1)$$

where P_i is the component of the total angular momentum along the i principal axis, μ_{ij} is the ij component of the inverse of the instantaneous inertia tensor and H_T and H_v represent the part of the Hamiltonian connected with the internal torsion and vibration, respectively. A Van Vleck transformation[34] similar to that used by Howard and Wilson[120] is applied to remove the nondiagonality in the vibrational quantum number v. The results may be expressed approximately as

$$H=W_v+\tfrac{1}{2}\sum_{ij}(v|\mu_{ij}|v)P_iP_j+(v|H_T|v)$$
$$+\sum_{v'}{}'|(v|\tfrac{1}{2}\sum_{ij}\mu_{ij}P_iP_j+H_T|v')|^2(W_v-W_{v'})^{-1}. \qquad (7\text{-}2)$$

The first term W_v, the vibrational energy, represents merely an additive constant and can be safely dropped. If only the splittings of the rotational lines in various torsional states are desired, then the terms independent of the rotational quantum numbers can be neglected. Finally, the terms corresponding to the ordinary centrifugal distortion effect can usually be ignored. By using Eq. (7-2) as the Hamiltonian, Kivelson[42] calculated the effective rotational constant B for a symmetric rotor and arrived at the equation

$$B_{v\sigma}=B_v+F_v(Kv\sigma|1-\cos3\alpha|Kv\sigma)+G_v(Kv\sigma|p'^2|Kv\sigma)$$
$$+L_vK(Kv\sigma|p'|Kv\sigma), \qquad (7\text{-}3)‡‡‡‡‡$$

where p' is the angular momentum associated with the torsional motion as defined in Eq. (2-8), B_v is the effective rotational constant averaged over the vibrational functions, and F_v, G_v, and L_v are empirical constants which depend on the interaction of the vibrations and the internal torsion. It is now possible to modify Eq. (7-3) slightly in order to see clearly the difference between this expression and the familiar one for ordinary vibrations. The torsional equation in reduced form is

$$\left[p'^2+\frac{9s}{8}(1-\cos3\alpha)\right]P_{Kv\sigma}=(9/4)b_{v\sigma}P_{Kv\sigma}. \qquad (7\text{-}4)$$

††††† Here the word "vibration" means all the internal modes of motion other than the hindered rotation.

‡‡‡‡‡ Kivelson[42] used the quantum number m for $Kv\sigma$ and Π_2 for $p'=p-\rho P_z$.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

In this equation the kinetic and potential energy may be identified

$$\langle T\rangle=(Kv\sigma|p'^2|Kv\sigma)$$

$$\langle V\rangle=\frac{9s}{8}(Kv\sigma|1-\cos 3\alpha|Kv\sigma).$$

For $K=0$ Eq. (7-3) becomes

$$B_{v\sigma}=B_v+F_v'\langle V\rangle+G_v\langle T\rangle, \qquad (7\text{-}5)$$

where

$$F_v'=8F_v/9s;$$

or

$$B_{v\sigma}=B_v-\beta_1\langle H\rangle-\beta_2[\langle T\rangle-\langle V\rangle], \qquad (7\text{-}6a)$$

$$=B_v-(9/4)\beta_1 b_{v\sigma}-\beta_2 x_{v\sigma}, \qquad (7\text{-}6b)$$

where

$$\beta_1=-\tfrac{1}{2}(G_v+F_v'), \quad x_{v\sigma}=\langle T\rangle-\langle V\rangle,$$
$$\beta_2=-\tfrac{1}{2}(G_v-F_v'), \quad \langle H\rangle=\langle T\rangle+\langle V\rangle=(9/4)b_{v\sigma}.$$

Now for harmonic vibrations

$$\langle T\rangle=\langle V\rangle,$$

and the asymptotic form for b_v as $s\rightarrow\infty$ is $b_v\sim 2(s)^{\frac{1}{2}}\times(v+\frac{1}{2})$. It follows that

$$B_{v\sigma}=B_v-\alpha_T(v+\tfrac{1}{2}), \qquad (7\text{-}7)^{29}$$

where

$$\alpha_T=(9/2)(s)^{\frac{1}{2}}\beta_1.$$

For the $J=1\leftarrow 0$ transition of CH_3SiH_3 Kivelson found

$$F_v=-65.322 \text{ Mc} \quad \text{and} \quad G_v=-0.8824 \text{ Mc},$$
$$s=30.22 \ (\alpha=17.0=\tfrac{9}{16}s),$$

and therefore $F_v'=-1.9212$ Mc, $\beta_1=+1.4018$ Mc, $\beta_2=-0.5194$ Mc, and $\alpha_T=+34.68$ Mc. For an ordinary vibration, the constant α is of the order B_e^2/ω which is ~ 22 Mc. Consequently, we see that this analysis is very similar to that for ordinary vibration rotation interaction; only the extreme departure from harmonicity must be corrected by the β_2 term.

Kivelson's method of analysis has also been extended to include asymmetric molecules[43,44,45b] and the Stark effect.[45a] Application of this theory has been made in the study of the vibration-torsion interaction of methyl alcohol,[77] methyl mercaptan[49,50a] and propylene oxide[27] (cf. Fig. 6).

Hecht and Dennison[21] have also investigated the nonrigidity effects in methyl alcohol. In addition to the centrifugal stretching, these authors point out that the effect of Coriolis force (from the internal motion) on the distortion of the equilibrium configuration of the molecule should also be considered. While the OH group executes the internal torsion, the Coriolis force arising from the end-over-end rotation of the entire molecule causes the hydrogen atom to rock back and forth on the COH plane. The effect is to produce a change in the average moments of inertia (over one external revolution) which, analogous to the distortion by centrifugal force described in the beginning of this section, depends on the torsional state. Hecht and Dennison[21] also take into account the fact that the height of the barrier is a function of the molecular dimension.[43] Inasmuch as the origin of the barrier is not clearly understood, the dependence of the barrier height on the displacements of the atoms cannot be calculated in any simple way; it can best be described in a parametric form as

$$V=V_3[1+\textstyle\sum_i\alpha_i\delta_{q_i}+\cdots] \qquad (7\text{-}8)$$

where the δ_{q_i} are the displacements of the atoms. A rigorous theoretical treatment which takes into consideration the effects of centrifugal and Coriolis forces on the average rotational constants as well as the change of potential barrier height could become very lengthy because of the large number of normal vibrations which could effect the moments of inertia through Coriolis interaction. Hecht and Dennison[21] carried out the calculation in two steps. They first examined the interaction between the rotation and the normal vibration in which the O—H bar moves in its own plane relative to the methyl group. In other words, they used a relatively simple model for CH_3OH which has five degrees of freedom—three for over-all rotation, one for internal torsion, and one for the O—H rocking. All the other modes of vibration were considered to be frozen. In the second step the remaining normal vibrations were introduced as perturbation and the effects of these additional terms on the energy levels were investigated with the aid of appropriate approximations. As the vibrational frequencies of methyl alcohol can be obtained from the vibrational spectrum, and the potential barrier height has been determined from the rotational transitions in the ground torsional states, the only parameter that needs to be introduced is the one which describes the variation of the potential barrier height with respect to the vibrational coordinate. Hecht and Dennison[21] have been able to explain quantitatively some thirty splittings of lines of normal methyl alcohol and five other isotopes using only six empirical parameters.

This treatment can, in principle, be applied to determine the potential barrier of symmetric top molecules for which the methods described previously (PAM and IAM) fail. However, the force constants for most of the symmetric top molecules are usually not known to a sufficient degree of accuracy and therefore it becomes necessary to introduce several empirically determined parameters. A reliable result on the barrier height is obtained only if a sufficiently large number of lines in the excited torsional levels can be observed.

Recently, Swan and Strandberg[103] reanalyzed the results on methyl alcohol considering vibration rotation interaction more completely. By using the observed

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

infrared vibrational spectrum for the normal frequencies they diagonalized the vibration-rotation-internal rotation matrix, i.e., they attempted to calculate the empirical constants F_v, G_v, etc., from the vibrational wave functions. As Hecht and Dennison point out, this procedure can, at present, only be used to verify constants but not calculate good values for them. This is borne out in Swan and Strandberg's results in which the quantitative fit of the calculated frequencies to the observed is only fair. Nevertheless, this type of calculation is valuable, and with improved wave function and energy values a better quantitative agreement on molecules might be obtainable.

II. Intensity Method

The ratio of the intensity (I_1/I_0) of given rotational transitions in the first excited torsional state to the ground torsional state is equal to the relative population in the two torsional states, i.e.,

$$I_1/I_0 = (g_1/g_0)e^{\Delta E_v/kT}, \qquad (7\text{-}9)$$

where g_0 and g_1 are the statistical weights of the ground and the excited torsional states and ΔE_v is the difference in energy between these two states. Since the ratio of g_1 and g_0 can be determined from group-theoretical consideration, a measurement of the relative intensities between a rotational line and its satellite gives the torsional frequency which, with the aid of the Mathieu tables,[36,104] leads to a determination of the potential barrier. Where there are uncertainties as to the ratio of the statistical weights of the two levels, the relative intensity can be measured at several different temperatures.

One of the first molecules studied by this technique was CH_3CF_3.[14,72] The barrier heights of a number of other molecules obtained by this procedure are summarized in Table X. The main virtue of the intensity method lies on its simplicity and its applicability to the case of very high barrier in which the splitting between the A and E lines are undetectably small. However, the experimental error involved in the relative intensity measurement is rather large—usually 10 to 20% or even higher. It is therefore not surprising that some of the resulting barrier heights obtained represent little more than estimates.

Baird and Bird[2] studied the problem of the measurement of relative intensities in the microwave region. They propose a procedure by which one can obtain an improvement in accuracy and reproducibility of the intensity ratio. Unfortunately, their technique or one comparable has generally not been used in the work reported in Table X. Recently, Verdier and Wilson[111] obtained more refined measurement on the relative intensities for the spectral lines of CH_3CHO and CH_3CH_2F using a Stark-modulated cavity. Their results on the barrier height compare favorably with those obtained by the "splitting methods" (see Table IX). They point out some limitations of the intensity method.

First, the quantity measured is usually the ratio of the peak absorption heights. The absorption coefficient of a line at its center for a Van Vleck-Weisskopf line shape is given in Eq. (7-10)[108]:

$$\alpha(\nu_0) = (16\pi^3 n_0 |\mu_{mn}|^2 \nu_0 \tau)/3kcT, \qquad (7\text{-}10)$$

where n_0 is the number of molecules per unit volume in the lower level of the transition, μ_{mn} is the matrix element of the dipole moment between the upper and the lower level, and τ is the mean interval between collisions. In order to compare the peak absorption coefficients of two lines ν_0, μ_{mn}, τ, and T must be the same, otherwise any difference will contribute to a correction term in the intensity ratio. The temperature T will undoubtedly be the same for the two lines. The correction for the difference in ν_0 can be made easily. However, the two are usually so nearly equal that this correction is negligible. The transition moment μ_{mn} may be different, because the asymmetry parameter could change from the ground to the excited state. For certain transitions in nearly symmetric tops, μ_{mn} may be quite sensitive to a small change in the asymmetry parameter and in such case an appropriate correction in Eq. (7-10) should be applied. The variation in τ between torsional levels should normally be within the limit of experimental accuracy.

The second difficulty which might arise is that the A and E lines are not resolved, and yet the separation between these lines is about the same order of magnitude as the line width; the measured peak intensity then does not correspond to Eq. (7-10). For example, it may happen that the A and E lines are resolved in the excited torsional state but overlap (with only one peak) in the ground state. In this case serious error would result if a correction were not applied to the measured intensity ratio. In order to correct for this effect, one can measure the integrated intensity of these lines instead of the peak value. The integrated intensity of the resolved A or E line is then one-half of that of the composite line. Alternatively, if the splitting of the unresolved doublet is calculated from theory, then a relation between the actual and apparent peak absorption can be derived by adding two Van Vleck-Weisskopf shaped lines separated by a frequency equal to the splitting of the doublet.

8. EXPERIMENTAL RESULTS

Table IX gives the experimental values for the barrier heights determined from frequency measurements of the microwave spectra. If more than one value is available in the literature, the best value according to our opinion is reported.§§§§§

§§§§§ The major references on each molecule are listed and the one which gave the best value for the barrier as shown in Table IX is underlined.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

TABLE IX. Barriers by the frequency method.

Compound	Formula	V (cm^{-1})	V (cal/mole)	Method	References
Threefold barriers					
Methyl alcohol	CH_3OH	374.8	1070	IAM	8, 32, 77
				PAM	98
				nonrigidity	21, 77, 103
				experimental	10, 12, 28, 30, 109, 110
Methyl amine	CH_3NH_2	691.1	1976	IAM	31, 78, 79
				experimental	28, 56–58 88–91
	CD_3NH_2	684.7 ± 2	1958	IAM	59, 46
Methyl mercaptan	CH_3SH	444 ± 10	1270	IAM	49, 50, 35
Acetic acid	CH_3COOH	169	483 ± 25	PAM	105
Acetaldehyde	CH_3CHO	406	1150 ± 30	PAM	38
Acetal fluoride	CH_3COF	378	1080	PAM	80
Acetal chloride	CH_3COCl	472	1350	PAM	93
Acetal cyanide	CH_3COCN	444	1270 ± 30	PAM	52
Methyl formate	$HCOOCH_3$	416	1190 ± 40	PAM	11b
Ethyl bromide	CH_3CH_2Br	1248	3567 ± 30	IAM 1st	66a
Ethyl chloride	CH_3CH_2Cl	1245	3560 ± 12	IAM 1st	66a
Ethyl fluoride	CH_3CH_2F	1158	3306	PAM 1st	22
1,1 difluoroethane	CH_3CHF_2	1112	3180	PAM 1st	22
Methyl silane	CH_3SiH_3	558 ± 17	1595	nonrigidity	42
		595	1700 ±100	PAM	39
Methyl fluorosilane	CH_3SiH_2F	545	1560	PAM	81a
Methyl difluorosilane	CH_3SiHF_2	439	1255	PAM	102
Methyl germane	CH_3GeH_3	433	1239 ± 25	PAM	53a
Methyl allene	$CH_3CH=C=CH_2$	556 ± 2	1589 ± 6	IAM	63
Propylene	$CH_3CH=CH_2$	692 ± 6	1978 ± 17	IAM	62
1-fluoropropylene	$CH_3CH=CHF$	752	2150 ±150	PAM	92
2-fluoropropylene	$CH_3CF=CH_2$	848	2424 ± 25	PAM	81c
Propylene oxide	$CH_3CH—CH_2$ (O)	895 ± 25	2560 ± 70	PAM IAM	101, 27
1-chlorobutyne 2	$CH_3C\equiv CCH_2Cl$	<30	<100	...	53b
Vinyl silane	CH_2CHSiH_3	523 ± 18	1495 ± 50	PAM	81d
Ethyl iodide	CH_3CH_2I	839	2400	PAM	33c
Sixfold barriers					
Nitromethane	CH_3NO_2	2.11	6.03± 0.03	low barrier	106, 107
	CD_3NO_2	1.82	5.19± 0.03	low barrier	107
Methyl boron difluoride	CH_3BF_2	4.82	13.77± 0.03	low barrier	75
Two threefold barriers					
Acetone	$(CH_3)_2CO$	266 ± 20	760	PAM	100
Dimethyl ether	$(CH_3)_2O$	950 ± 50	2720	PAM	33b
Dimethyl silane	$(CH_3)_2SiH_2$	582	1665	PAM	81b
Twofold barriers					
Hydrogen peroxide	H_2O_2	113	323	IAM	70, 71
Phenol	C_6H_5OH	1100 ±100	3140	...	50b

As discussed in Secs. 3, 4, and 5, the ratio V_3/F is determined from the analysis of a microwave spectrum. In general, this ratio can be determined to an accuracy of two percent or better. Uncertainties in the molecular structure and limitations in the approximations are the two main sources of error. The spectra of molecules with lower s values are influenced to a much larger extent by the internal rotation and usually these can be analyzed to give somewhat more accurate values for the ratio V_3/F. Structural errors, in regard to the reduced moment of inertia of the methyl group, cause a large share of the error in the barrier. In our opinion, the barrier heights reported in Table IX are accurate to somewhere between 2% and 5%.

The spectra of methyl alcohol and methyl amine have been studied very thoroughly in the microwave region. Furthermore, a considerable amount of work has been done on their theoretical interpretation using

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

the IAM given in Sec. 3. To a lesser extent methyl mercaptan has also been studied. These three molecules are of the class of an intermediate barrier molecule with slight asymmetry. Mainly the perpendicular ($\Delta K = \pm 1$) Q branch lines ($\Delta J = 0$) have been analyzed for the barrier height. The effects of the asymmetry and the off-diagonal internal rotational matrix elements are calculated by second-order perturbation. Methyl amine has the additional complication of inversion similar to ammonia.

Also the $J = 1 \leftarrow 0$, $K = 0$ transition has been analyzed for methyl alcohol[21,77,103] including the effects of nonrigidity. The splittings of the internal rotation doublets are caused by the rigid internal torsion as well as the molecular vibration, but the division between these two effects is not known accurately since it depends rather critically on the structure of the molecule. Slight shifts in the methyl group axis affect this division. With six empirical constants a good agreement between the theory and experiment was obtained for the frequency shifts of the excited states transitions for a number of isotopes of methyl alcohol. Nishikawa[77] obtained a good fit of the spectrum with only the three symmetric top constants. Attempts have been made[103] to calculate these "empirical constants" from the vibrational wave functions and qualitative agreement between the calculated and observed values has been obtained.

The results of the barrier height of methyl alcohol and methyl amine obtained by various methods are all in good agreement. However, for methyl mercaptan the situation is somewhat confusing. Intensity work by Solimene and Dailey[95] gave a value of 371 cm^{-1}. Kilb[35] determined a value of 246 cm^{-1} by rigid hindered theory (PAM to second order) on the $J = 1 \leftarrow 0$ transitions of a large number of isotopic species measured by Solimene and Dailey.[95] Kojima and Nishikawa first obtained a barrier[49] of 280 cm^{-1}, but subsequently they revised their value[50] to 444 cm^{-1}. Also the $J = 1 \leftarrow 0$ lines of the isotopic species $C^{12}H_3SH$, $C^{13}H_3SH$, $C^{12}D_3SH$, and $C^{12}H_3SD$ were analyzed by the symmetric top nonrigidity theory. According to Kojima and Nishikawa the $J = 1 \leftarrow 0$ of $C^{12}H_3SD$ was incorrectly assigned by Solimene and Dailey. The results of Kojima and Nishikawa[50] are probably more accurate because their analysis was carried out to a higher order of approximation.

The molecules acetaldehyde (CH_3CHO), acetyl fluoride (CH_3COF), acetyl chloride (CH_3COCl), acetyl cyanide (CH_3COCN), propylene ($CH_3CH=CH_2$), 1 fluoropropene ($CH_3CH=CHF$), methyl allene ($CH_3CH=C=CH_2$), methyl silane (CH_3SiH_3), methyl fluorosilane (CH_3SiFH_2), and methyl difluorosilane (CH_3SiF_2H) are examples of molecules with relatively high barriers. A pseudo-rigid rotor splitting was used to determine the barrier height. In some cases transitions up to $J = 12$ were analyzed using the PAM. Also in the near symmetric top molecules the first-order (nonrigid rotor) terms produced a large effect. This was especially true in acetaldehyde[38] (see Sec. 5) and methyl silane, mono-, and di-deuterated.[39] Methyl silane has been analyzed by nonrigidity theory[42] and all of the determined barrier heights agree very well with one another.

The barrier heights in ethyl fluoride,[22] 1,1-difluoroethane,[22] ethyl chloride,[66a] and ethyl bromide[66a] were determined by the splittings in the first excited torsional state since the ground state splittings were too small to observe. To a good approximation the splittings of the satellites of the first excited torsional state are caused entirely by the hindered rotation (but not the molecular vibrations) because of the high barrier.

In propylene oxide[27] the barrier was determined by the splittings in the ground and first two excited torsional states. The agreement is exceedingly good indicating the validity of using excited state splittings for the determination of barrier heights in molecules of high barriers. No effects of a V_6 potential were observed.

Acetic acid[105] is a molecule with an intermediate barrier and large asymmetry. Unusual difficulty was encountered in the analysis, but, by using the PAM to fourth order, Tabor was able to assign a number of higher J lines. The assignment was one of the more difficult aspects of this problem. The IAM would have been very difficult to use in this particular case because of the relatively high J transitions necessary for its analysis.

Low barriers have been observed in nitromethane and methyl boron difluoride. Here the $m = 3$ lines give an accurate determination of the barrier height. In the low barrier problem the initial assignment is quite difficult and an extensive searching of the lines is needed. The accuracy of these barrier heights is exceedingly good and generally better than that in the high barrier case. This results from the wide splittings of the order of 1000 Mc as compared with small splittings of a few Mc of the high barrier problems. The very low barrier obtained in these two cases for a sixfold potential seem to indicate that the ratio of V_3 to V_6 is generally of the order of 100. 1-chloro butyne-2 is another molecule with a low barrier which has been recently studied.[53b] Only an upper limit of the barrier height has been obtained.

The barriers have been determined for a number of molecules by the intensity method of Sec. 7. These are listed in Table X. Generally, an error of 10 to 20% can by assigned to these determined barrier heights. It is interesting to compare some of the results in Table X with the more accurate values of Table IX. Ethyl bromide[115,116] is 21% low; ethyl chloride[112,114], 4.5% low; ethyl fluoride[111] is in agreement; and 1,1-difluoroethane[94] is 12% high. By taking into account a 5% error in the values in Table IX, we can see that 10–20% error in the values of Table X is quite realistic.

It is interesting to compare the results of the barrier heights obtained by microwave spectroscopy with those from other means, particularly the thermodynamic methods (see Sec. 1). While detailed examinations of

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

TABLE X. Barriers by the intensity method.

Compound	Formula	V (cm^{-1})	V (cal/mole)	References
Acetaldehyde	CH_3CHO	386	1103± 60	111
Ethyl bromide	CH_3CH_2Br	980	2800±500	115, 116
Ethyl chloride	CH_3CH_2Cl	1190	3400±600	112, 114
Ethyl cyanide	CH_3CH_2CN	1150±100	3280±290	54
Ethyl fluoride	CH_3CH_2F	1160	3310±210	111, 51
1,1 Difluoroethane	CH_3CHF_2	1250±200	3570±580	94
1,1,1 Trifluoroethane	CH_3CF_3	1200	3480	14, 72
Methyl mercaptan	CH_3SH	371± 43	1060±120	95
Methyl silane	CH_3SiH_3	460± 80	1315	60
Methyl trifluorosilane	CH_3SiF_3	420	1200	72, 73, 87
1 Trifluoro butyne-2	$CH_3-C\equiv C-CF_3$	<100	<300	3, 4a, 61
Isobutane	$(CH_3)_3CH$	1360	3900	65
t-Butyl fluoride	$(CH_3)_3CF$	1500	4300	65
Trimethyl phosphine	$(CH_3)_3P$	910	2600	65
Trimethyl amine	$(CH_3)_3N$	1530	4400	64
Methyl germane	CH_3GeH_3	585	1673	4b

the thermodynamic results are beyond the scope of this paper, it may be stated that the barrier heights determined by the splitting methods are usually more accurate than those obtained from the thermodynamic properties of the molecules. On the other hand, the intensity and thermodynamical methods yield results of comparable accuracy (ten to twenty percent uncertainty). With careful experimental measurements and in favorable cases, this error may be somewhat lower for both of these methods. In the absence of any frequency work on a given molecule, a careful investigation of the specific experimental difficulties is required in order to decide whether the thermodynamic results or the microwave intensity results are more reliable.

In addition to the barrier heights, the equilibrium configuration has been obtained for a number of molecules listed in Table XI. These equilibrium configurations are valuable because they supply additional information which, besides the barrier heights, any theory of the origin of barriers must explain. Ethyl chloride monodeuterated[113] was the first molecule in which the staggered configuration was confirmed by microwave spectroscopy. The tunneling rate is much slower than the measurement time in the microwave region so the two different configurations are observed: the deuterium gauche and trans to the chlorine atom. Each of these was observed. Even with only an approximate structure these observed rotational constants would be inconsistent with an eclipsed configuration.

Finally, a few molecules have been studied recently with two internal rotors. Although the theory, an extension of the work of Sec. 3, is not covered here [see references 84, 40, 100, and 33(b)], the results on acetone, dimethyl ether, and dimethyl silane are given in Table IX. The theory of internal rotation for molecules with an asymmetric internal rotor is not covered in this review, but two molecules, hydrogen peroxide,[70] and phenol,[50b] have been studied and analyzed approximately as a symmetric top. References 6, 9, and 86 give some of the theoretical work.

TABLE XI. Equilibrium configuration.

Compound	Formula	Configuration	Reference
Acetaldehyde	CH_2DCHO CHD_2CHO	Methyl group eclipses O and staggers H	38
Acetyl chloride	CH_2DCOCl CHD_2COCl	Methyl group eclipses O and staggers Cl	93
Acetyl cyanide	CH_2DCOCN CHD_2COCN	Methyl group eclipses O and staggers CN	52
Acetal fluoride	CH_2DCOF CHD_2COF	Methyl group eclipses O and staggers F	80
Propylene	$CH_2DCH=CH_2$	Methyl group eclipses the double bond and staggers H	26
Ethyl chloride	CH_2D-CH_2Cl	staggered	113
Methyl silane	CH_2D-SiH_2D $CH_2D-SiHD_2$	staggered	39
Methyl fluorosilane	CH_2D-SiH_2F	staggered	81a
1,1-Difluorosilane	CH_2DSiHF_2 CHD_2SiHF_2	staggered	25c

9. CONCLUSIONS

After the various discussions on the interactions of internal rotation and over-all rotation, one naturally questions what causes these potential barriers and whether the values reported in Sec. 8 can be explained quantitatively in terms of the electronic structure of the molecules. Furthermore, can unknown barrier heights be predicted satisfactorily? Numerous attempts have been made (see Bibliography—Theoretical and Empirical Methods of Barrier Height Estimation) to provide answers to these questions, but the results do not seem very encouraging. This is understandable since the potential barrier represents only a very small fraction of the binding energy of the molecule. It seems highly impractical at this time to make any serious attempt to evaluate the barrier heights by the direct

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

and rigorous approach, i.e., from the difference in energy of the entire molecule for the two extreme configurations corresponding to the eclipsed form ($\alpha=60°$) and the staggered form ($\alpha=0°$).

In order to offer a simple picture for the nature of the hindering potential, various mechanisms such as Van der Waals forces,[012,016] direct electrostatic interactions,[03,010,013] overlap, and exchange force,[014] have been proposed for the intramolecular interaction. Recently, Wilson[017,018] has made a critical examination of these various hypotheses and found that none of the existing theories can account for the experimental results of the barrier heights of the various molecules. He [017] suggests that the potential barrier could be an inherent property of the axial bond itself and that it is is not due to the direct forces between the attached atoms to any great extent or to the electronic distribution in the attached bonds which are at any considerable distance. Pauling,[014] on the other hand, hypothesizes that exchange interactions of the electrons in the attached bonds, as determined by the overlap of these attached bonds which extend in the direction of the axial bond toward one another, are responsible for the potential barrier. By the use of higher orbitals, d and f, he makes a rough estimate of this nonbonding interaction which is in the right order of magnitude for the observed potential barriers. Although the absolute values are only fortuitous, a number of interesting relative values are presented. From his hypothesis, unshared electron pairs, which do not hybridize with the f orbitals, would not contribute to the potential barrier, i.e., methyl alcohol would have a barrier approximately one-third of that in ethane and methyl amine would have a barrier approximately two-thirds of that of ethane. Another interesting point is the existence of the very low barriers in molecules such as nitromethane. A sixfold barrier would require i orbitals with an azimuthal quantum number $l=6$. Because of the large promotional energy these probably would not contribute much to the hybrid bond orbitals and hence we would predict a low barrier.

Which of these various mechanisms, or combination of mechanisms, correctly accounts for the potential barriers measured in molecules is still unanswered. It is hoped that this review will stimulate further research toward the understanding of the origin of the potential energy hindering internal rotation. Nevertheless, from the viewpoint of molecular dynamics, the problem of internal rotation in molecules, i.e., the effect of the internal rotation on the energy levels, is quite well understood. The agreement between the theory and the observed spectra, including all the fine details, is particularly outstanding. Furthermore, the study of the microwave spectra offers a method, which is by far the most accurate at the present time, for determining the height and the shape of the hindering barrier.

APPENDIX 1

Derivation of Kinetic Energy and Angular Momentum

In Fig. 9, O and C represent, respectively, the center of mass of the entire molecule and of the top alone. Two vectors $\mathbf{r}_j$ and $\boldsymbol{\sigma}_j$ are drawn to a typical atom in the top from the respective centers of mass O and C. The position vector (with O as origin) for an atom in the framework will be denoted by $\mathbf{r}_i$. If $\boldsymbol{\varrho}$ denotes the vector from O to C, one can readily see that

$$\mathbf{r}_j=\boldsymbol{\varrho}+\boldsymbol{\sigma}_j. \tag{A1-1}$$

The kinetic energy can be written as

$$T=\tfrac{1}{2}\sum_i m_i(\partial\mathbf{r}_i/\partial t)\cdot(\partial\mathbf{r}_i/\partial t) +\tfrac{1}{2}\sum_j m_j(\partial\mathbf{r}_j/\partial t)\cdot(\partial\mathbf{r}_j/\partial t). \tag{A1-2}$$

By remembering that

$$(\partial\mathbf{r}_i/\partial t)=\boldsymbol{\omega}\times\mathbf{r}_i,\quad (\partial\mathbf{r}_j/\partial t)=\boldsymbol{\omega}\times\mathbf{r}_j+(\partial\boldsymbol{\alpha}/\partial t)\times\boldsymbol{\sigma}_j,$$

where ω is the angular velocity of the over-all rotation of the molecule, one can rewrite (A1-2) in the form of

$$\begin{aligned}T&=\tfrac{1}{2}\sum_i m_i(\boldsymbol{\omega}\times\mathbf{r}_i)^2+\tfrac{1}{2}\sum_j m_j[(\boldsymbol{\omega}\times\mathbf{r}_j)+(\partial\boldsymbol{\alpha}/\partial t\times\boldsymbol{\sigma}_j)]^2\\ &=\tfrac{1}{2}\sum_i m_i[r_i^2\omega^2-(\boldsymbol{\omega}\cdot\mathbf{r}_i)^2]+\tfrac{1}{2}\sum_j m_j[r_j^2\omega^2-(\boldsymbol{\omega}\cdot\mathbf{r}_j)^2]\\ &\quad+\tfrac{1}{2}\sum_j m_j[\sigma_j^2\dot{\alpha}^2-(\boldsymbol{\sigma}_j\cdot\partial\boldsymbol{\alpha}/\partial t)^2]\\ &\quad+\sum_j m_j(\boldsymbol{\omega}\times\mathbf{r}_j)\cdot(\partial\boldsymbol{\alpha}/\partial t\times\boldsymbol{\sigma}_j).\end{aligned} \tag{A1-3}$$

The first two terms obviously represent the kinetic energy of the over-all rigid rotation of the molecule with angular velocity $\boldsymbol{\omega}$, and can be written as $\frac{1}{2}\boldsymbol{\omega}^+\cdot\mathbf{I}\cdot\boldsymbol{\omega}$ where $\mathbf{I}$ is the usual inertia tensor of the entire molecule. The third term may be expressed as $\frac{1}{2}I_\alpha\dot{\alpha}^2$ where I_α is the moment of inertia of the top about its own axis of symmetry. The last term may be simplified as

$$\begin{aligned}&\sum_j m_j(\boldsymbol{\omega}\times\mathbf{r}_j)\cdot(\partial\boldsymbol{\alpha}/\partial t\times\boldsymbol{\sigma}_j)\\ &=\sum_j m_j\{[(\boldsymbol{\varrho}+\boldsymbol{\sigma}_j)\cdot\boldsymbol{\sigma}_j](\boldsymbol{\omega}\cdot\partial\boldsymbol{\alpha}/\partial t)\\ &-(\boldsymbol{\omega}\cdot\boldsymbol{\sigma}_j)[(\boldsymbol{\varrho}+\boldsymbol{\sigma}_j)(\partial\boldsymbol{\alpha}/\partial t)]\}\\ &=\sum_j m_j[\sigma_j^2(\boldsymbol{\omega}\cdot\partial\boldsymbol{\alpha}/\partial t)-(\boldsymbol{\omega}\cdot\boldsymbol{\sigma}_j)(\boldsymbol{\sigma}_j\cdot\partial\boldsymbol{\alpha}/\partial t)].\end{aligned} \tag{A1-4}$$

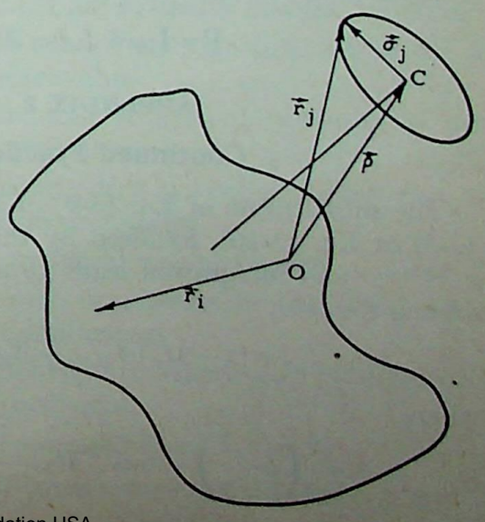


FIG. 9. The position vectors of typical atoms in a molecule with an attached internal rotor. The O position and C position represent the centers of mass of the whole molecule and the internal rotor, respectively. The vectors $\mathbf{r}_i$ and $\mathbf{r}_j$ are vectors from the center of mass O to typical atoms in the frame of the molecule and the internal rotor.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

The last step follows from the relation

$$\sum_j m_j \boldsymbol{\sigma}_j = 0.$$

By using a coordinate system attached to the top with the z axis along the axis of symmetry, and the origin located at the center of mass of the top, one has

$$\boldsymbol{\sigma}_j = x_j\hat{\imath} + y_j\hat{\jmath} + z_j\hat{k},$$

and

$$\sum_j m_j x_j z_j = \sum_j m_j y_j z_j = 0.$$

Equation (A1-4) then takes the form of

$$\begin{aligned}&\sum_j m_j[(x_j^2+y_j^2+z_j^2)(\boldsymbol{\omega}\cdot\partial\boldsymbol{\alpha}/\partial t)\\&\quad-(\omega_x x_j+\omega_y y_j+\omega_z z_j)z_j|\partial\boldsymbol{\alpha}/\partial t|]\\&=\sum_j m_j[(x_j^2+y_j^2+z_j^2)(\boldsymbol{\omega}\cdot\partial\boldsymbol{\alpha}/\partial t)-\omega_z z_j^2|\partial\boldsymbol{\alpha}/\partial t|]\\&=\sum_j m_j(x_j^2+y_j^2)(\boldsymbol{\omega}\cdot\partial\boldsymbol{\alpha}/\partial t)=I_\alpha(\boldsymbol{\omega}\cdot\partial\boldsymbol{\alpha}/\partial t).\end{aligned}$$

The kinetic energy finally becomes

$$T=\tfrac{1}{2}\boldsymbol{\omega}^+\cdot\mathbf{I}\cdot\boldsymbol{\omega}+\tfrac{1}{2}I_\alpha\dot{\alpha}^2+I_\alpha(\boldsymbol{\omega}\cdot\partial\boldsymbol{\alpha}/\partial t). \quad \text{(A1-5)}$$

The classical angular momentum is given by

$$\begin{aligned}\mathbf{P}&=\sum_i m_i(\mathbf{r}_i\times\partial\mathbf{r}_i/\partial t)+\sum_j m_j(\mathbf{r}_j\times\partial\mathbf{r}_j/\partial t)\\&=\sum_i m_i[\mathbf{r}_i\times(\boldsymbol{\omega}\times\mathbf{r}_i)]\\&\quad+\sum_j m_j[\mathbf{r}_j\times\{(\boldsymbol{\omega}\times\mathbf{r}_j)+(\partial\boldsymbol{\alpha}/\partial t\times\boldsymbol{\sigma}_j)\}]\\&=\sum_i m_i[\mathbf{r}_i\times(\boldsymbol{\omega}\times\mathbf{r}_i)]+\sum_j m_j[\mathbf{r}_j\times(\boldsymbol{\omega}\times\mathbf{r}_j)]\\&\quad+\sum_j m_j[\mathbf{r}_j\times(\partial\boldsymbol{\alpha}/\partial t\times\boldsymbol{\sigma}_j)]. \quad \text{(A1-6)}\end{aligned}$$

The first two terms stand for the angular momentum of the over-all rigid rotation of the entire molecule which can be written as $\mathbf{I}\cdot\boldsymbol{\omega}$. In order to simplify the last term, use is made of the fact that $\boldsymbol{\sigma}_j$ is a vector measured from the center of mass of the top and that the axis of the top is a principal axis of the top. It follows that

$$\begin{aligned}&\sum m_j[\mathbf{r}_j\times(\partial\boldsymbol{\alpha}/\partial t\times\boldsymbol{\sigma}_j)]\\&\quad=\sum m_j[(\boldsymbol{\varrho}+\boldsymbol{\sigma}_j)\times(\partial\boldsymbol{\alpha}/\partial t\times\boldsymbol{\sigma}_j)]\\&\quad=\sum m_j[\boldsymbol{\sigma}_j\times(\partial\boldsymbol{\alpha}/\partial t\times\boldsymbol{\sigma}_j)]\\&\quad=\sum m_j[\sigma_j^2\partial\boldsymbol{\alpha}/\partial t-(\boldsymbol{\sigma}_j\cdot\partial\boldsymbol{\alpha}/\partial t)\boldsymbol{\sigma}_j]\\&\quad=\sum m_j[\sigma_j^2\hat{k}|\partial\boldsymbol{\alpha}/\partial t|-z_j^2\hat{k}|\partial\boldsymbol{\alpha}/\partial t|]=I_\alpha\partial\boldsymbol{\alpha}/\partial t.\end{aligned}$$

Equation (A1-6) then reduces to

$$\mathbf{P}=\mathbf{I}\cdot\boldsymbol{\omega}+I_\alpha\partial\boldsymbol{\alpha}/\partial t.$$

APPENDIX 2

Continued Fraction

The substitution of Eq. (3-9) or Eq. (3-23) into Eq. (3-3) or Eq. (3-16), followed by the multiplication by $e^{-i(3k+\sigma)\alpha}$ and integration leads to a recursion formula among the A's:

$$A_{3(k-1)+\sigma}+(\lambda-M_k)A_{3k+\sigma}+A_{3(k+1)+\sigma}=0, \quad \text{(A2-1)}$$

where

$$\lambda=\frac{4}{s}\left(b-\frac{s}{2}\right) \quad \text{and} \quad M_k=\frac{16}{9s}(3k+\sigma)^2.$$

This defines an infinite set of homogeneous equations. In order to have a nontrivial solution it is necessary that the determinant of the coefficients vanish. This infinite determinant can be factored into 3 different subdeterminants, one for each σ value. Koehler and Dennison[48] have shown that Eq. (A2-1) leads to a continued fraction. This can easily be seen by dividing the recursion relation Eq. (A2-1) by $A_{3k+\sigma}$:

$$\frac{A_{3(k+1)+\sigma}}{A_{3k+\sigma}}=(M_k-\lambda)-\frac{A_{3(k-1)+\sigma}}{A_{3k+\sigma}}, \quad \text{(A2-2)}$$

and inverting

$$\frac{A_{3k+\sigma}}{A_{3(k+1)+\sigma}}=\frac{1}{M_k-\lambda-\dfrac{A_{3(k-1)+\sigma}}{A_{3k+\sigma}}}=\frac{1}{M_k-\lambda-\dfrac{1}{M_{k-1}-\lambda-\cdots}}=G_k^-,$$

also

$$\frac{A_{3k+\sigma}}{A_{3(k-1)+\sigma}}=\frac{1}{M_k-\lambda-\dfrac{1}{M_{k+1}-\lambda-\cdots}}=G_k^+.$$

By substituting back into Eq. (A2-1), one obtains a continued fraction equation,

$$\lambda=M_k-G_{k+1}^+-G_{k-1}^-. \quad \text{(A2-3)}$$

A trial value of λ is substituted into G^+ and G^- leading to a first approximation of λ. By successive iteration a consistent value of the eigenvalue, λ, is obtained. The continued fraction does not always converge very rapidly and can even diverge. The use of a Newton Raphson approximation[69] is recommended to avoid this difficulty. By letting

$$f(\lambda)=M_k-\lambda-G_{k+1}^+(\lambda)-G_{k-1}^-(\lambda)=0,$$

and expanding at the nth approximation of the eigenvalue (denoted by λ^n), one obtains

$$\begin{aligned}f(\lambda)&=f(\lambda^n)+(\lambda-\lambda^n)f'(\lambda^n)\\&=\lambda^{n+1}-\lambda^n+(\lambda-\lambda^n)f'(\lambda^n)=0.\end{aligned}$$

Then

$$\lambda=\lambda^n+\frac{\lambda^{n+1}-\lambda^n}{-f'(\lambda^n)}, \quad \text{(A2-4)}$$

where

$$\begin{aligned}-f'(\lambda^n)&=1+G_{k+1}^{+2}[1+G_{k+2}^{+2}(1+\cdots)]\\&\quad+G_{k-1}^{-2}[1+G_{k-2}^{-2}(1+\cdots)].\end{aligned}$$

Since the G's are determined in the calculation of the eigenvalue, the derivative $f'(\lambda^n)$ is readily calculated. Care must also be taken to insure that the initial approximation is not separated from the desired eigenvalue by a pole. A pole occurs between each eigen-

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

value. If one exists between the first approximation and the desired root, the continued fraction converges to the adjacent root.

With the desired eigenvalue determined, a substitution into Eq. (A2-1) leads to a recursion relation between the A's. By solving these relations and normalizing, one obtains the eigenfunctions. These eigenfunctions are used to determine the matrix elements and integrals for the IAM and PAM if the approximations given are not sufficiently accurate. The eigenfunctions are tabulated for the periodic solutions.[104,37] For the nonperiodic solutions σ is replaced by $\sigma+\rho K$ in all the previous expressions.

APPENDIX 3

Tabulations Related to Mathieu Equations

The most recent and complete tabulations are listed in Table XII. References 36 and 104 are the most useful and complete for the threefold potential case. Also, by the use of Eq. (3-29) and these tables, the eigenvalues for $s>30$ for other periodicities can readily be determined. A number of older references are given in reference 104.

TABLE XII. Mathieu function tabulations.

Reference	Period in x	Range in V	Range in s	Accuracy
Mathieu Functions[a,c]	π, 2π	0(1)8 9(1)14	0(1)100[b] 0(2)100	1×10^{-8}
Kilb[c,d,e]	3π	0(1)5	2(1)30(2)52(4)100	1×10^{-10}
Stejskal[f]	π,3π	0(1)4	100(15)205	1×10^{-7}
Blanch and Rhodes[g]	π, 2π	0(1)15	100[b]	1×10^{-8}
Pexton[i]	4π	0(1)9	1(1)40	1×10^{-6}

[a] See reference 104.
[b] A close tabulation is given for the low s values.
[c] Eigenfunctions are also tabulated.
[d] Copies are available from E. B. Wilson, Jr., Department of Chemistry, Harvard University. Herschbach's tables are also available.
[e] See reference 36.
[f] See reference 96.
[g] See reference 5.
[h] Tabulated in terms of t 0(0.002)0.1 where $s=1/t^2$.
[i] Tabulated in Li and Pitzer reference 55. Notation: $4\theta=s$, $\theta_0=b-\frac{1}{2}s$.

Recently, Herschbach[23] tabulated the matrix elements p, p^2, p^3, cos6α, the perturbation sums to fourth order for use in the PAM, and the Fourier coefficients for use in the IAM.[25b] Herschbach gives a complete introduction explaining the use of the tables to solve the internal rotation problem. These tables are of great assistance in the computation for barrier problems.

APPENDIX 4

Notation

a	IAM "a" coordinate axis
a, a_K	expansion coefficients
a, a'	low barrier diagonal elements $\lvert m\rvert=3$
a_k	Fourier coefficient in potential energy
a_n	Fourier coefficient in IAM internal energy
amu	atomic mass units
A	symmetry species in C_3 group
A, A_a, $A_{v\sigma}$, A_x	rotational constants
$A_{3k+\sigma}{}^{v}$, $A_{3k+\sigma}{}^{Kv}$	wave function expansion coefficients
A	Angstrom unit
b	IAM "b" coordinate axis
b	expansion coefficients
b	low barrier off-diagonal element $\lvert m\rvert=3$
b, be, bo, b_i, $b(N\pi)$	Mathieu eigenvalues
B, B_b, $B_{v\sigma}$, B_y	rotational constants
B_e	equilibrium rotational constants
B, B_x, B_y, B_z	symmetry species
c	IAM "c" coordinate axis
c_k	Fourier coefficient of Mathieu's function
C, C_c, $C_{v\sigma}$, C_z	rotational constants
C_2, C_3, C_6, C_x	rotational symmetry operations
d	reducing factor, r, times $(I_xI_yI_z)$
d_k	Fourier coefficient of Mathieu's function
d_{XY}	interatomic distance from atom X to atom Y
D_{bc}, D_{ij}	Hamiltonian cross terms between components of the total angular momentum
De, Do	Fourier coefficients of Mathieu's function
$D^J(\beta)_{KpK}$	rotational transformation matrix element
E	symmetry species, degenerate
E	energy
$E(\kappa)$	reduced rigid rotor energy
E_{JKM}	rotational energy
$E_{v\sigma}$, $E_{Kv\sigma}$	torsional energy: PAM and IAM
$E(JKMv\sigma)$	total energy: rotation and torsion
f	Mathieu parameter (Floquet's theorem)
$f(x)$	IAM approximate integral function
$f(\lambda)$	continued fraction function of the λ eigenvalue
F	reduced rotational constant of the internal rotor
F_v	nonrigidity potential energy expansion coefficient
g_0, g_1	statistical weight factors
g^+, g^-	correction terms in the PAM rotational energy
G_v	nonrigidity kinetic energy expansion coefficient
$G_k{}^+$, $G_k{}^-$	continued fraction terms
$\hbar$	Planck's constant divided by 2π

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

H, H_0	Hamiltonian
H_{IR}	coupling term in Hamiltonian
H_R	rotational Hamiltonian
H_T	torsional Hamiltonian
H_v	vibration Hamiltonian, harmonic oscillator function
H_1	asymmetry term in Hamiltonian
$H_{v\sigma}$	rotational Hamiltonian for the $v\sigma$ torsion state
$\hat{\imath}$	unit vector
$\mathbf{I}$	moment of inertia tensor
$I_a, I_b, I_c, I_{aa}, I_{bb}$ $I_{cc}, I_g, I_x, I_y, I_z, I_{jk}$	moments of inertia
I_α	moment of inertia of the internal rotor about its symmetry axis
I_{bc}, I_{jk}	products of inertia
$\hat{\jmath}$	unit vector
J	quantum number of the total angular momentum
k	Boltzmann's constant
$\hat{k}$	unit vector
k^+, k^-	correction terms in the PAM rotational energy
K	quantum number of the total angular momentum along the molecular z axis
l	free rotation quantum number
L_v	nonrigidity angular momentum expansion coefficient
m	free rotation quantum number
m_i	mass of the ith atom
M	quantum number of the total angular momentum along the space axis
$M_{Kv\sigma}(\alpha)$	IAM nonperiodic torsional function
$M(x)$	Mathieu function
Mc	megacycles/second
M_k	continued fraction term
N	periodicity of the hindering potential
p	internal angular momentum
$p_{v\sigma,v'\sigma}$	PAM matrix element of the internal angular momentum
$P_g(P_a,P_b,P_c,P_j,P_k,P_x,P_y,P_z)$	the gth components of the angular momentum
P_g^{ij}	the gth component of the total angular momentum with the independent variables i and j constant
$P_{Kv\sigma}(\alpha)$	IAM torsional function for a symmetric molecule
Q_i	PAM Hamiltonian term
$Q_{Kv\sigma}(\alpha)$	IAM torsional function for an asymmetric molecule
r	reducing factor
$\mathbf{r}_i, \mathbf{r}_j$	position vector of the ith (j) atom
$r_a, r_b, r_c, r_x, r_y, r_z$	position vector components
s	reduced barrier height
S_1	rotational transformation matrix
S_2	asymmetric transformation matirx
Se, So	Mathieu functions
$S_{JKM}(\theta,\varphi)e^{iK\chi}$	symmetric rotor function
T	kinetic energy
T	temperature
$U_{v\sigma}(\alpha)$	PAM torsional function
v	torsional quantum number
$V(\alpha), V_N, V_3, V_6, V_{12}$	potential energy (subscript denotes periodicity)
$w_{v\sigma}^{(n)}$	PAM perturbation terms
W, W_r	rigid rotor energy
x	x principal axis
$x_{v\sigma}$	nonrigidity term
y	y principal axis
z	z principal axis
α	internal angle
α	direction cosine of ϱ to z principal axis
$\dot{\alpha}$	internal angular velocity
α_t, α_i	first term in the vibration rotation expansion of the rotational constant
β	direction cosine of ϱ to y principal axis
β_1, β_2	nonrigidity expansion terms
Δ_0	IAM splitting of internal sublevels for $K=0$
Δ_K	IAM splitting of internal sublevels for Kth level
ΔE_{JK}	internal rotational splitting of the J, K rotational level
$\Delta E(J_{K-1K+1})$	internal rotational splitting of the J_{K-1K+1} rotational level
δ_{Asy}	asymmetry splitting of K and $-K$ rotational levels
δ_K	nonrigid-rotor internal rotor splitting
$\delta_{K'v'\sigma}{}^{Kv\sigma}$	IAM integrals
δq_i	displacement coordinate
θ	Eulerian angle
κ	asymmetry parameter
λ	eigenvalue of a secular determinant
$\lambda_g, (\lambda_x,\lambda_y,\lambda_z)$	direction cosine of the internal rotor to the g principal axis
μ_{mn}	dipole matrix element
ν	frequency
Π	dot product of the total angular momentum vector and the vector $\mathbf{p}$

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

ϱ	(a) direction of the internal axis such that the internal angular momentum vanishes	σ_j	position vector of a top atom from the center of mass of the top (internal rotor)
	(b) position vector of the center of mass of the top from the center of mass of the whole molecule (Appendix 1)	τ	mean life time
		φ, Φ	Eulerian angles
		$\chi, \chi_1, \chi_2, \chi'$	Eulerian angles
		Ψ, ψ	wave function
		ω	total angular velocity vector
ρ_g	component of ϱ along the g axis	ω_g ($\omega_a, \omega_b, \omega_c, \omega_x, \omega_y, \omega_z$)	component of the total angular velocity along the g axis
σ	torsional sublevel quantum number		

APPENDIX 5

Comparison of the Notation of Various Authors (Grouped Where Possible)

Present notation	[a]	[b]	[c]	[d]
I_g, I_{ij}	A, B, C, C_1, D	I_{xx}, I_{yy}, I_{zz}	A, B, C, C_1, I_{jk}	I_x, I_y, I_z
I_α	C_2	I_1	C_2	I_α
ρ	$(C_2/C)^*$	...	λ	(ρ)
r	$\frac{(BC_1-D^2)}{(BC-D^2)}$	...	$\frac{(BC_1-D^2)}{(BC-D^2)}$	r, d
σ	τ^e	...	μ	$\sigma(A,E), \kappa$
m	k	k_1	...	m
$w_{v\sigma}{}^{(1)}$	...	...	...	$-2p_{vv}$
$w_{v\sigma}{}^{(2)}$	...	...	...	$1+4F\rho$
$E_{Kv\sigma}, E_{v\sigma}$	$E_{K\tau n}$	...	$E^{K\mu n}$	$E_{v\sigma}$
$\Delta E(J_{K-1K+1})$	$(J_{K-1K+1}), K$	...	...	$\Delta A, \Delta W$
P_g	P_g	J_g	P_g	P_g
p	p_x	L_1	P	p
α	x	φ	x	α
ρ_y/ρ	$\sin\beta\sim\beta$	...	λ_η/λ	...
s	H'	...	α	s
$M_{Kv\sigma}(\alpha), P_{Kv\sigma}(\alpha), Q_{Kv\sigma}(\alpha), U_{v\sigma}(\alpha)$	$P_{K\tau n}(x)$	...	$Q^{K\mu n}$	$U_{v\sigma}(\alpha)$
v	n	...	n	v

[a] Dennison, Koehler, Burkhard, Ivash, Hecht, Lide, and Mann.
[b] Gwinn, Myers, and Tannenbaum.
[c] Itoh.
[d] Wilson, Lin, Lide, Kilb, Swalen, and Herschbach.
[e] $[K+\tau-\sigma=1 \pmod 3]$.

ACKNOWLEDGMENTS

It is a pleasure to acknowledge the helpful comments and suggestions given by a large number of our friends and associates. We are particularly indebted to E. B. Wilson, Jr., D. R. Herschbach, R. W. Kilb, V. W. Laurie, L. C. Krisher, and T. N. Sarachman for their criticism and encouragement on this paper. We also wish to express our appreciation to D. G. Burkhard, A. E. Douglas, K. T. Hecht, H. E. Hoffman, T. Kojima, K. S. Pitzer, R. M. St. John, and M. W. P. Strandberg for many helpful discussions and correspondences concerning various parts of this review. Our thanks are also extended to Mrs. J. E. Locke, Mrs. V. S. Blohm, and particularly Mrs. Mary Swalen for typing the manuscript.

BIBLIOGRAPHY

1 E. R. Andrew, *Nuclear Magnetic Resonance* (Cambridge University Press, New York, 1955).
2 D. H. Baird and G. R. Bird, Rev. Sci. Instr. **25**, 319 (1954).
3 Bak, Christensen, Hansen-Nygaard, and Tannenbaum, J. Chem. Phys. **26**, 241 (1957).
4a Bak, Hansen, and Anderson, J. Chem. Phys. **21**, 1612 (1953).
4b A. I. Barchukov and A. M. Prokhorov, Optika i Septroskopia **4**, 799 (1958).
5 G. Blanch and I. Rhodes, Wash. Acad. Sci. **45**, 166 (1955).
6 D. G. Burkhard, J. Chem. Phys. **21**, 1541 (1953).

SUBJECT[a]

T	NMR
TX	Intensity measurements
X	$CH_3C{\equiv}CCF_3$
X	$CH_3C{\equiv}CCF_3$
IX	CH_3GeH_3
T	Mathieu, high s
T	Asym.-asym.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

[7]	D. G. Burkhard, Trans. Faraday Soc. **52**, 1 (1956).	T	IAM
[8]	D. G. Burkhard and D. M. Dennison, Phys. Rev. **84**, 408 (1951).	T	CH_3OH IAM
[9]	D. G. Burkhard and J. C. Irvin, J. Chem. Phys. **23**, 1355 (1955).	T	Asym.-asym.
[10]	D. K. Coles, Phys. Rev. **74**, 1194 (1948).	X	CH_3OH
[11a]	B. L. Crawford, J. Chem. Phys. **8**, 273 (1940).	T	PAM
[11b]	R. F. Curl, Jr. J. Chem. Phys. **30**, 1529 (1959).	X	$HCOOCH_3$ PAM
[12]	B. P. Dailey, Phys. Rev. **72**, 84 (1947).	X	CH_3OH
[13]	B. P. Dailey, Ann. N. Y. Acad. Sci. **55**, 915 (1952).	Review	
[14]	Dailey, Shulman, and Minden, Phys. Rev. **75**, 1319A (1949).	XI	CH_3CF_3
[15]	T. Das, J. Chem. Phys. **25**, 896 (1956).	T	Mathieu
[16]	T. Das, J. Chem. Phys. **27**, 763 (1957).	T	Mathieu
[17]	H. S. Gutowsky and C. H. Holm, J. Chem. Phys. **25**, 1228 (1956).	TX	NMR amides
[18]	J. O. Halford, J. Chem. Phys. **18**, 444 (1950).	T	Mathieu
[19]	J. O. Halford, J. Chem. Phys. **26**, 851 (1957).		
[20]	K. T. Hecht and D. M. Dennison, J. Chem. Phys. **26**, 31 (1957).	T	IAM
[21]	K. T. Hecht and D. M. Dennison, J. Chem. Phys. **26**, 48 (1957).	T	CH_3OH nonrigidity
[22]	D. R. Herschbach, J. Chem. Phys. **25**, 358 (1956).	TX	Excited states CH_3CH_2F and CH_3CHF_2
[23]	D. R. Herschbach, "Tables for the Internal Rotation Problem," Department of Chemistry, Harvard University (1957).	T	Mathieu
[24]	D. R. Herschbach, J. Chem. Phys. **27**, 975 (1957).	T	Mathieu
[25a]	D. R. Herschbach, J. Chem. Phys. **27**, 1420 (1957).	T	IAM-PAM connection
[25b]	D. R. Herschbach, J. Chem. Phys. **31**, 91 (1959).	T	PAM theory
[25c]	D. R. Herschbach (private communications).	TX	
[26]	D. R. Herschbach and L. C. Krisher, J. Chem. Phys. **28**, 728 (1958).	X	Equilibrium configuration CH_2DCHCH_2
[27]	D. R. Herschbach and J. D. Swalen, J. Chem. Phys. **29**, 761 (1958).	TX	$CH_3CH{-}CH_2$ PAM (O)
[28]	W. D. Herschberger and J. Turkevich, Phys. Rev. **71**, 554 (1947).	X	CH_3OH and CH_3NH_2
[29]	G. Herzberg, *Infrared and Raman Spectra* (D. Van Nostrand Company, Inc., Princeton, New Jersey, 1945).	T	Vibration and rotation spectra
[30]	Hughes, Good, and Coles, Phys. Rev. **84**, 418 (1951).	X	CH_3OH
[31]	T. Itoh, J. Phys. Soc. Japan **11**, 264 (1956).	T	CH_3NH_2 IAM
[32]	E. V. Ivash and D. M. Dennison, J. Chem. Phys. **21**, 1804 (1953).	T	CH_3OH IAM
[33a]	Ivash, Li, and Pitzer, J. Chem. Phys. **23**, 1814 (1955).	T	CH_3OH
[33b]	P. H. Kasai and R. J. Myers, J. Chem. Phys. **30**, 1096 (1959).	TX	$(CH_3)_2O$
[33c]	T. Kasuya and T. Oka, J. Phys. Soc. Japan **14**, 980 (1959).	X	C_2H_5I
[34]	E. C. Kemble, *Fundamental Principles of Quantum Mechanics* (McGraw-Hill Book Company, Inc., New York, 1937), p. 394.	T	Van Vleck transformation
[35]	R. W. Kilb, J. Chem. Phys. **23**, 1736 (1955).	T	CH_3SH PAM
[36]	R. W. Kilb, "Tables of Mathieu Eigenvalues and Eigenfunctions Mathieu for Special Boundary Conditions," Department of Chemistry, Harvard University (1956).	T	Mathieu
[37]	R. W. Kilb, thesis, Harvard University (1956).	T	Asym.-asym.
[38]	Kilb, Lin, and Wilson, J. Chem. Phys. **26**, 1695 (1957).	TX	CH_3CHO PAM
[39]	R. W. Kilb and L. Pierce, J. Chem. Phys. **27**, 108 (1957).	TX	CH_3SiH_3 PAM
[40]	J. E. Kilpatrick and K. S. Pitzer, J. Chem. Phys. **17**, 1064 (1949).	T	Compound rotation
[41]	King, Hainer, and Cross, J. Chem. Phys. **11**, 27 (1943).	T	Asym. rotor
[42]	D. Kivelson, J. Chem. Phys. **22**, 1733 (1954).	T	CH_3SiH_3 nonrigidity
[43]	D. Kivelson, J. Chem. Phys. **23**, 2230 (1955).	T	Nonrigidity
[44]	D. Kivelson, J. Chem. Phys. **23**, 2236 (1955).	T	Nonrigidity
[45a]	D. Kivelson, J. Chem. Phys. **26**, 215 (1956).	T	Stark effect
[45b]	D. Kivelson, J. Chem. Phys. **27**, 980 (1957).	T	Errata
[46]	D. Kivelson and D. R. Lide, Jr., J. Chem. Phys. **27**, 353 (1957).	T	CD_3ND_2 IAM
[47]	D. Kivelson and E. B. Wilson, Jr., J. Chem. Phys. **20**, 1575 (1952).	T	Centrifugal distortion
[48]	J. S. Kohler and D. M. Dennison, Phys. Rev. **57**, 1006 (1940).	T	IAM
[49]	T. Kojima and T. Nishikawa, J. Phys. Soc.-Japan **10**, 240 (1955).	TX	CH_3SH IAM
[50a]	T. Kojima and T. Nishikawa, J. Phys. Soc. Japan, **12**, 680 (1957).	TX	CH_3SH IAM
[50b]	T. Kojima (private communication).	X	C_6H_5OH
[51]	J. Kraitchman and D. P. Dailey, J. Chem. Phys. **23**, 184 (1955).	X	CH_3CH_2F intensity
[52]	L. C. Krisher and E. B. Wilson, Jr. (to be published).	TX	CH_3COCN PAM
[53a]	V. W. Laurie, J. Chem. Phys. **30**, 1210 (1959).	X	CH_3GeH_3 PAM
[53b]	V. W. Laurie and D. R. Lide, Jr., Bull. Am. Phys. Soc. Ser. II, **3**, 415 (1958).	X	$CH_3C{\equiv}CCH_2Cl$
[54]	R. G. Lerner and B. P. Dailey, J. Chem. Phys. **26**, 678 (1957).	IX	CH_3CH_2CN
[55]	J. C. M. Li and K. S. Pitzer, J. Phys. Chem. **60**, 466 (1956).	T	IAM Mathieu
[56]	D. R. Lide, Jr., J. Chem. Phys. **20**, 1812 (1952).	X	CH_3NH_2
[57]	D. R. Lide, Jr., J. Chem. Phys. **21**, 571 (1953).	X	CH_3NH_2
[58]	D. R. Lide, Jr., J. Chem. Phys. **22**, 1613 (1954).	X	CH_3NH_2
[59]	D. R. Lide, Jr., J. Chem. Phys. **27**, 343 (1957).	TX	CD_3ND_2 IAM
[60]	D. R. Lide, Jr., and D. K. Coles, Phys. Rev. **80**, 911 (1950).	X	CH_3SiH_3 intensity
[61]	D. R. Lide, Jr., and D. Kivelson, J. Chem. Phys. **23**, 2191 (1955).	T	CH_3CCCF_3
[62]	D. R. Lide, Jr., and D. E. Mann, J. Chem. Phys. **27**, 868 (1957).	TX	CH_3CHCH_2 IAM
[63]	D. R. Lide, Jr., and D. E. Mann, J. Chem. Phys. **27**, 874 (1957).	TX	CH_3CHCCH_2 IAM
[64]	D. R. Lide, Jr., and D. E. Mann, J. Chem. Phys. **28**, 572 (1958).	TIX	$(CH_3)_3N$
[65]	D. R. Lide, Jr., and D. E. Mann, J. Chem. Phys. **29**, 914 (1958).	XI	$(CH_3)_3X$ (X = CH, CF, and P)
[66a]	D. R. Lide, Jr., J. Chem. Phys. **30**, 37 (1959).	X	Excited states IAM CH_3CH_2Cl, CH_3CH_2Br

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

[66b]	D. R. Lide, Jr. (to be published).	X	$(CH_3)_3As$
[67]	C. C. Lin, Amer. J. Phys. **26**, 319 (1958).	T	Classical internal rotation
[68]	C. C. Lin and R. W. Kilb, J. Chem. Phys. **24**, 631 (1956).	TX	CH_3CHO PAM
[69]	H. Margenau and G. M. Murphy, *Mathematics of Physics and Chemistry* (D. Van Nostrand Company, Inc., Princeton, New Jersey, 1943).	T	Newton Raphson
[70]	J. T. Massey and D. R. Bianco, J. Chem. Phys. **22**, 442 (1954).	TX	H_2O_2
[71]	J. T. Massey and R. W. Hart, J. Chem. Phys. **23**, 942 (1956).	T	H_2O_2
[72]	H. Minden and B. P. Dailey, Phys. Rev. **82**, 338A (1951).	XI	CH_3CF_3, CH_3SiF_3
[73]	Minden, Mays, and Dailey, Phys. Rev. **78**, 347A (1950).	XI	CH_3SiF_3
[74]	R. E. Naylor, Jr., thesis, Harvard University (1955).	TX	CH_3BF_2
[75]	R. E. Naylor, Jr. and E. B. Wilson, Jr., J. Chem. Phys. **26**, 1057 (1957).	TX	CH_3BF_2
[76]	H. H. Nielsen, Phys. Rev. **40**, 445 (1932).	T	IAM sym.
[77]	T. Nishikawa, J. Phys. Soc. Japan **11**, 781 (1956).	T	CH_3OH IAM nonrigidity
[78]	T. Nishikawa, J. Phys. Soc. Japan **12**, 668 (1957).	TX	CH_3NH_2 IAM
[79]	Nishikawa, Itoh, and Shimoda, J. Chem. Phys. **23**, 1735 (1955).	X	CH_3NH_2
[80]	L. Pierce and L. C. Krisher (to be published).	TX	CH_3COF PAM
[81a]	L. Pierce, J. Chem. Phys. **29**, 383 (1958).	TX	CH_3SiH_2F PAM
[81b]	L. Pierce, J. Chem. Phys. **31**, 547 (1959).	TX	$(CH_3)_2SiH_2$
[81c]	L. Pierce and J. O'Reilly (to be published).	X	CH_3CFCH_2 PAM
[81d]	L. Pierce and J. O'Reilly (private communication).	X	CH_2CHSiH_3 PAM
[82]	K. S. Pitzer, J. Chem. Phys. **14**, 239 (1946).	T	Asym.-asym.
[83]	K. S. Pitzer, Discussions Faraday Soc. **10**, 66 (1951).	Review	
[84]	K. S. Pitzer and W. D. Gwinn, J. Chem. Phys. **10**, 428 (1940).	T	IAM
[85]	W. D. Phillips, J. Chem. Phys. **23**, 1363 (1955).	NMR	$CH_3CON(CH_3)_2$
[86]	Donna Price, J. Chem. Phys. **9**, 807 (1941); **10**, 80 (1942).	T	Asym.-asym.
[87]	J. Sheridan and W. Gordy, J. Chem. Phys. **19**, 965 (1951).	XI	CH_3SiF_3
[88]	K. Shimoda and T. Nishikawa, J. Phys. Soc. Japan **8**, 133 (1953).	X	CH_3NH_2
[89]	K. Shimoda and T. Nishikawa, J. Phys. Soc. Japan **8**, 425 (1953).	X	CH_3NH_2
[90]	Shimoda, Nishikawa, and Itoh, J. Phys. Soc. Japan **9**, 974 (1954).	X	CH_3NH_2
[91]	Shimoda, Nishikawa, and Itoh, J. Chem. Phys. **22**, 1456 (1954).	X	CH_3NH_2
[92]	S. Siegal, J. Chem. Phys. **27**, 989 (1957).	X	CH_3CHCHF PAM
[93]	K. Sinnott, Bull. Am. Phys. Soc. Ser. II **1**, 198 (1956).	TX	CH_3COCl PAM
[94]	N. Solimene and B. P. Dailey, J. Chem. Phys. **22**, 2042 (1954).	X	CH_3CHF_2 intensity
[95]	N. Solimene and B. P. Dailey, J. Chem. Phys. **23**, 124 (1955).	X	CH_3SH intensity
[96]	E. O. Stejskal and H. S. Gutowsky, J. Chem. Phys. **28**, 388 (1958).	T	Mathieu and NMR
[97]	M. W. P. Strandberg, *Microwave Spectroscopy* (Metheun and Company, Ltd., London, 1954).	T	Microwave spectroscopy
[98]	J. D. Swalen, J. Chem. Phys. **23**, 1739 (1955).	T	CH_3OH PAM
[99]	J. D. Swalen, J. Chem. Phys. **24**, 1072 (1956).	T	Approx formula PAM
[100]	J. D. Swalen and C. C. Costain (to be published).	TX	$(CH_3)_2CO$ PAM
[101]	J. D. Swalen and D. R. Herschbach, J. Chem. Phys. **27**, 100 (1957).	TX	CH_3CH-CH_2 PAM (O)
[102]	J. D. Swalen and B. P. Stoicheff, J. Chem. Phys. **28**, 671 (1958).	TX	CH_3SiHF_2 PAM
[103]	P. R. Swan and M. W. P. Strandberg, J. Mol. Spectroscopy **1**, 33 (1957).	T	CH_3OH
[104]	*Tables Relating to Mathieu Functions* (Columbia University Press, New York, 1951).	T	Mathieu
[105]	W. J. Tabor, J. Chem. Phys. **27**, 974 (1957).	TX	CH_3COOH PAM
[106]	Tannenbaum, Johnson, Myers, and Gwinn, J. Chem. Phys. **22**, 949 (1954).	TX	CH_3NO_2
[107]	Tannenbaum, Myers, and Gwinn, J. Chem. Phys. **25**, 42 (1956).	TX	CH_3NO_2
[108]	C. H. Townes and A. L. Schawlow, *Microwave Spectroscopy* (McGraw-Hill Book Company, Inc., New York, 1955).	T	Microwave spectroscopy
[109]	Venkateswarlu, Edwards, and Gordy, J. Chem. Phys. **23**, 1195 (1955).	X	CH_3OH
[110]	P. Venkateswarlu and W. Gordy, J. Chem. Phys. **23**, 1200 (1955).	X	CH_3OH
[111]	P. H. Verdier and E. B. Wilson, Jr., J. Chem. Phys. **29**, 340 (1958).	TXI	Cavity spectrometer CH_3CHO, CH_3CH_2F
[112]	R. S. Wagner and B. P. Dailey, J. Chem. Phys. **22**, 1459 (1954).	XI	CH_3CH_2Cl
[113]	R. S. Wagner and B. P. Dailey, J. Chem. Phys. **23**, 1355 (1955).	X	CH_2DCH_2Cl configuration
[114]	R. S. Wagner and B. P. Dailey, J. Chem. Phys. **26**, 1588 (1957).	XI	CH_3CH_2Cl
[115]	Wagner, Dailey, an Solimene, J. Chem. Phys. **26**, 1593 (1957).	XI	CH_3CH_2Br
[116]	Wagner, Solimene, and Dailey, J. Chem. Phys. **23**, 599 (1955).	XI	CH_3CH_2Br
[117]	E. Wigner, *Gruppentheorie* (Vieweg and Sohn, Braunschweig, 1931).	T	Group theory
[118]	E. B. Wilson, Jr., J. Chem. Phys. **3**, 276 (1935).	T	Symmetry
[119]	E. B. Wilson, Jr., Chem. Revs. **27**, 17 (1940).	T	PAM Mathieu
[120]	E. B. Wilson, Jr. and J. B. Howard, J. Chem. Phys. **4**, 260 (1936).	T	Vibration rotation interaction
[121]	Wilson, Lin, and Lide, J. Chem. Phys. **23**, 136 (1955).	T	PAM

[a] T, theoretical paper; X, experimental paper; I, intensity method.

Bibliography: Theoretical and Empirical Methods of Barrier Height Estimation

[01] Aston, Discussions Faraday Soc. **10**, 72 (1952).
[02] Aston, Isserow, Szasz, and Kennedy, J. Chem. Phys. **12**, 336 (1944).
[03] Au Chin Tang, J. Chinese Chem. Soc. **18**, 1 (1951); Sci. Sinica **3**, 279 (1954).
[04] B. Bak, J. Chem. Phys. **24**, 918 (1956).
[05] H. Eyring, J. Am. Chem. Soc. **54**, 3191 (1932).
[06] Eyring, Stewart, and Smith, Proc. Natl. Acad. Sci. U. S. 44, 259 (1958).
[07] French and Rasmussen, J. Chem. Phys. **14**, 389 (1946).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

[08] Gorin, Walter, and Eyring, J. Am. Chem. Soc. **61**, 1876 (1939).
[09] R. W. Kilb, "Internal rotation and microwave spectroscopy," thesis, Harvard University (1956).
[010] Lasettre and Dean, J. Chem. Phys. **17**, 317 (1949); **16**, 151 (1948).
[011] Luft, Discussions Faraday Soc. **10**, 117 (1951); J. Chem. Phys. **21**, 179 (1953); *ibid.* **22**, 155, 1814 (1954); Z. Elektrochem. **59**, 46 (1955); Trans. Faraday Soc. **49**, 118 (1953).
[012] E. A. Mason and M. M. Kreevoy, J. Am. Chem. Soc. **77**, 5808 (1955).
[013] Oosterhoff, Discussions Faraday Soc. **10**, 79 (1951).
[014] L. Pauling, Proc. Natl. Acad. Sci. U. S. **44**, 211 (1958).
[015] K. Pitzer, Discussions Faraday Soc. **10**, 66 (1951).
[016] J. Van Dranen, J. Chem. Phys. **20**, 1982 (1952).
[017] E. B. Wilson, Jr., Proc. Natl. Acad. Sci. U. S. **43**, 816 (1957).
[018] E. B. Wilson, Jr., "The Problem of Barriers to Internal Rotation in Molecules," *Advances in Chemical Physics* (Interscience Publishers, Inc., New York, 1959), Vol. II.
[019] W. Weltner, Jr., J. Chem. Phys. **31**, 264 (1959).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

Foundations of the Optical Model for Nuclei and Direct Interaction

G. E. BROWN

Department of Mathematical Physics, University of Birmingham, Brimingham, England

I. INTRODUCTION

THIS paper gives a description of the scattering of particles of energies in the kev and Mev range by complex nuclei. Starting from exact many-level formulas for the scattering amplitude, we show that several phenomena which were inexplicable in the old compound-nucleus theory arise from phase relations between the nuclear levels, whereas terms in which phase relations would not seem to be important correspond to the scattering predicted by the old compound-nucleus picture, supplemented by statistical assumptions.

This paper might be described as a treatment of the theory of nuclear reactions in the region in which many nuclear levels participate in the scattering. Thus, it is complementary to the comprehensive article of Lane and Thomas[1] where the treatment is particularly appropriate to the case in which few levels contribute. These authors employed the R-matrix theory of Wigner and Eisenbud[2] which is especially well adapted to the low-energy case because it makes the energy dependence of all expressions as explicit as possible. We employ here the formalism of Kapur and Peierls[3] which has the great advantage for the many-level case that the sum over levels enters linearly into the scattering amplitude. The formalism employed is only an intermediate step to the final results, and, therefore, we choose the one that seems simplest for the particular development here.

Surveying first the old compound-nucleus picture, we note the points at which it must be revised. It was realized long ago that this was not a complete description; in fact N. Bohr[4] already suggested in 1938 that a particle, upon entering the nucleus, might go directly to a final state without forming a compound nucleus. However, most calculations were carried out with the extreme form of the model, and most physical pictures tended to follow in this way. By extreme form we mean the form in which it is assumed that the phases of contributions from different compound levels are random, and that they therefore do not interfere. Such assumptions are generally implicit in statistical calculations.

The compound-nucleus picture formulated by N. Bohr[5] explained the very narrow resonances observed in the elastic scattering of slow neutrons by assuming that the incident neutron, once inside the nucleus, shares its energy with many other nucleons through its strong interaction with them; the resulting compound nucleus then lasts a long time, until one of the nucleons acquires sufficient energy to escape from the nucleus. Because of the long lifetime of the state, the uncertainty principle shows that its energy can be well determined, and hence, that its width is small. The observed resonances were of the order of electron volts wide, about a million times narrower than the single-particle levels formed in the scattering of particles by the potential wells earlier assumed to represent the nucleus.

Since the observed compound-nucleus resonances were so long-lived, it was usually assumed that their characteristics are independent of the way in which they are formed, aside from restrictions resulting from conservation laws, and are also independent of the neighboring states. It seemed natural then to introduce statistical assumptions which neglected interference between the scattering from different levels. Then, in the neighborhood of a compound state, one could write the cross section $\sigma_{\alpha\beta,j}$ for a particle incident in "channel" α to emerge in "channel" β (we give precise definitions of all terms later), leaving the residual nucleus in excited state j as a product of factors

$$\sigma_{\alpha\beta,j}=\frac{\pi}{k^2}(2l+1)\frac{\gamma_{p\alpha}\gamma_{p\beta j}}{(\epsilon_p-E)^2+\alpha_p^2/4},$$

where p labels the compound state, $\gamma_{p\alpha}$ is the partial width[6] for formation of it by a particle in channel α, $\gamma_{p\beta j}$ is the partial width for its decay, α_p is the total width of the state, and ϵ_p is the real part of its energy. This type of formula, valid in the region of a single isolated level, was given by Breit and Wigner[7] and bears their names. The above expression for σ as a product of factors implies the independence of the processes of formation and decay of the level both from each other and from the properties of other levels.

It is clear that the number of compound states per unit energy interval increases rapidly with the energy of the incident particle. Already at energies in the kev range, the density of levels observed in the scattering of slow neutrons by medium and heavy nuclei is $\sim 10^4$

[1] A. M. Lane and R. G. Thomas, Revs. Modern Phys. **30**, 257 (1958).
[2] E. P. Wigner and L. Eisenbud, Phys. Rev. **72**, 29 (1947).
[3] P. L. Kapur and R. E. Peierls, Proc. Roy. Soc. (London) **A166**, 277 (1938).
[4] N. Bohr, Nature **141**, 326, 1096 (1938); Bohr, Peierls, and Placzek, Nature **144**, 200 (1939).
[5] N. Bohr, Nature **137**, 344 (1936); Science **86**, 161 (1947).
[6] Usually the widths are denoted by capital letters Γ, e.g., $\Gamma_{p\alpha}$. However, we reserve these for the widths of the single-particle levels in a potential well.
[7] G. Breit and E. P. Wigner, Phys. Rev. **49**, 579 (1936).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

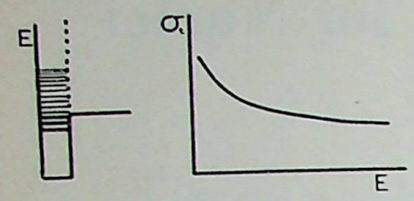


FIG. 1. The increasing density of nuclear levels with excitation energy is sketched on the left. Predictions of the continuum model for the behavior of the total cross section are shown on the right.

to 10^6 levels per Mev. Therefore, in the scattering of beams of neutrons which generally includes a spread in neutron energies of several Mev because of the experimental difficulty of producing monoenergetic neutrons of arbitrary energy, one might expect the cross section to approach that of a completely absorbing system with increasing neutron energy. A specific model incorporating these features was formulated by Feshbach, Peaslee, and Weisskopf[8] and is known as the continuum model. Here, the wave function of the neutron incident on the nucleus is subjected to the requirement that it have only an incoming part at the nuclear radius, i.e., that

$$d/dr[r\psi(r)]|_{r=R} = -iKR\psi(R),$$

where $\psi(r)$ is the wave function for the incident neutron, R is the nuclear radius, and K is an "internal" wave number that the neutron is supposed to possess inside the nucleus as a result of its interaction with the other nucleons. The predictions of such a theory for the total cross section are illustrated in Fig. 1 which also gives an indication of how the number of compound states increases with increasing excitation energy. The negative energy states of the nucleus are shown schematically as bound states in a potential well. At high energies the total cross section approaches $2\pi R^2$. The rise in the curve at lower energies is due to the fact that, in this model, the particle "feels" the nucleus already at distance $R+\lambda\!\!\!^{-}$, where $\lambda\!\!\!^{-}=\lambda/2\pi$ and λ is the wavelength, because of quantum-mechanical effects.

Contrary to expectations, the experimental cross sections, for beams in which the neutrons had a spread of energies, showed giant maxima of widths of the order of one or two Mev, as indicated schematically in Fig. 2. These were reproduced theoretically in the later theory of Feshbach, Porter, and Weisskopf[9] by the scattering from a complex potential well in which the real part represents some average interaction of the incoming nucleon with the nuclear particles, and the imaginary part, the disappearance of particles out of the incident

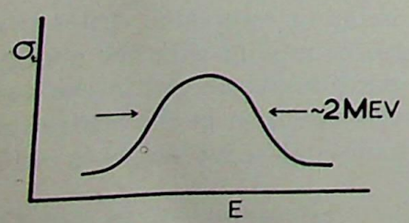


FIG. 2. Sketch of a giant resonance of the type observed in the scattering of neutrons by complex nuclei.

beam into compound states. There is no natural explanation for such giant resonances in the extreme compound-nucleus picture, and the representation of the nucleus by a complex well indicates a major modification.

Other phenomena which demanded modification of the extreme compound-nucleus picture were observed in the inelastic scattering of nucleons by complex nuclei. The predictions of the compound-nucleus model, supplemented by statistical assumptions, are shown as the solid line in Fig. 3. In this picture, the probability of leaving the nucleus in the various excited levels is assumed the same for all levels, aside from kinematical factors. The curve therefore rises as the energy taken off by the scattered particle decreases (and the energy left to the nucleus increases) and is cut off only when the wave number of the escaping nucleon (or nucleons) is so small that it will be reflected at the nuclear surface. Experimentally, an anomalously large number of fast particles was observed, as indicated by the dotted line in Fig. 3. Processes responsible for the high-energy particles were qualitatively well described by the direct interaction formalism of Austern, Butler, and

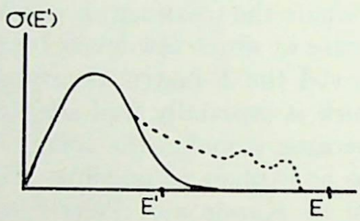


FIG. 3. The solid line shows typical predictions of the statistical model for the spectrum of inelastically-scattered neutrons. The energy of the emitted neutron is denoted by E'. The dashed line indicates the type of spectrum observed experimentally.

McManus.[10] Here, the incident particle is assumed to "chip" off particles from the nuclear surface without formation of a compound state. For example, the transition element for the nucleus to go from state i to state f is assumed to be

$$M_{if} = -\frac{2M}{\hbar^2}\int_{r_0}^{\infty} \chi_f(\xi) e^{-i\mathbf{k}_f\cdot\mathbf{r}} V(\mathbf{r}-\xi) e^{i\mathbf{k}_i\cdot\mathbf{r}} \chi_i(\xi) d^3\xi d^3r,$$

where $\exp i\mathbf{k}_i\cdot\mathbf{r}$ and $\exp i\mathbf{k}_f\cdot\mathbf{r}$ are the wave functions of the incident and scattered nucleon, and $\chi_0(\xi)$ and $\chi_f(\xi)$ those of the initial and residual nucleus. Here $\mathbf{r}$ represents the coordinate of the incident nucleon and ξ the totality of nuclear coordinates, conventions that are used throughout. The radius r_0 is supposed to define the point at which the "surface" begins and was chosen so as to fit the experimental angular distributions.

The form of this theory was suggested by the earlier description by Butler[11] of deuteron stripping which is also a form of "direct interaction." However, for sim-

[8] Feshbach, Peaslee, and Weisskopf, Phys. Rev. **71**, 145 (1947).
[9] Feshbach, Porter, and Weisskopf, Phys. Rev. **96**, 448 (1954).
[10] Austern, Butler, and McManus, Phys. Rev. **92**, 350 (1953).
[11] S. T. Butler, Proc. Roy. Soc. (London) **A208**, 559 (1951).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

plicity, we discuss the scattering of nucleons since complications are introduced by the composite structure of deuterons or α-particles.

Many experiments have shown that the high-energy particles in these inelastic processes have angular distributions, usually peaked towards small angles, whereas the low-energy nucleons have comparatively structureless, symmetrical angular distributions, as would be predicted by the statistical theory, and this is a further indication of the direct nature of the high-energy processes.

A further phenomenon which violates the early compound-nucleus picture occurs in the dipole photoeffect. Here, the absorption cross section for all complex nuclei shows a gaint maximum in the region of 15 Mev for heavy nuclei. This maximum was explained by Goldhaber and Teller[12] and Steinwedel and Jensen[13] in terms of a collective oscillation of proton density. However, the decay of such a system, in which the energy is almost evenly distributed over all particles, ought to be adequately represented by the statistical theory. In the case of medium and heavy nuclei, the number of high-energy protons observed is far in excess of that predicted by this theory, often by several orders of magnitude. Such high-energy protons are predicted naturally by Wilkinson's[14] picture of the giant dipole photoeffect, in which the γ-ray is absorbed by a nucleon which then goes into a single-particle state in a complex well, and consequently has an appreciable chance of escaping with the full energy of the γ-ray before being absorbed into compound states. (This absorption is again described through the imaginary part of the well.)

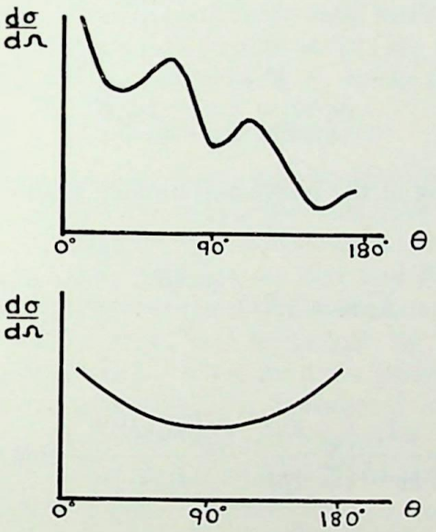


FIG. 4. In the upper figure, the angular distribution of the high-energy part of the inelastic spectrum, Fig. 3, is sketched; in the lower, that of the low-energy part of the spectrum.

Thus, although the old compound-nucleus picture had to be modified, the new phenomena could be reproduced by simple physical models which employed a complex well. However, conceptually, the description was not complete because it was known from cases where the cross sections could be investigated in detail that the giant resonances actually consist of thousands or millions of compound levels, and the relation between the detailed behavior—which often could not be investigated experimentally, but could be inferred—and the average behavior as predicted by the complex well had to be clarified. Further, the connection of the parameters of the optical well with more fundamental quantities such as the nucleon-nucleon force had to be made.

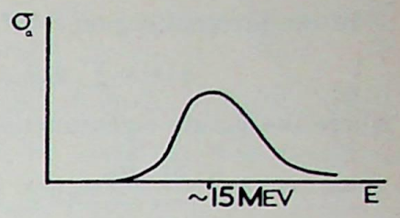


FIG. 5. The absorption cross section for γ-rays showing the giant dipole resonance.

A unified description of the above phenomena begins from an exact description in terms of nuclear dispersion theory. The scattering amplitude is separated into terms corresponding to direct interactions and compound-nucleus processes. Before going into such a description, however, we review the dispersion theory of Kapur and Peierls,[3] which is used in the later development, and present a simplified picture illustrating the physical assumptions employed later in the more formal arguments.

II. KAPUR-PEIERLS DISPERSION FORMALISM

1. Scattering by a Potential Well

We begin by treating the scattering of an S-wave neutron by a potential well; this simple example illustrates the main features of the theory. The Schrödinger equation for this case is

$$\frac{h^2}{2M}\frac{d^2\phi}{dr^2}+[E-V(r)]\phi=0, \tag{1}$$

where $V(r)$ is the potential and $\phi(r)=r\psi(r)$ with ψ the wave function of the neutron. For radii r greater than some radius R, beyond which the potential $V(r)$ is zero,

$$(d^2\phi/dr^2)+k^2\phi=0, \quad r>R, \tag{2}$$

with

$$k^2=2ME/\hbar^2. \tag{2.1}$$

The solution of Eq. (2) is

$$\phi=(\sin kr/k)+Se^{ikr}, \tag{2.2}$$

where the normalization of ϕ has been chosen so that the first term on the right-hand side corresponds to the S-wave part of a plane wave of unit amplitude, i.e., $\exp i\mathbf{kr}\cong\sin kr/kr$ for $kr\ll 1$; S is the amplitude of the scattered (outgoing) wave.

[12] M. Goldhaber and E. Teller, Phys. Rev. 74, 1046 (1948).
[13] H. Steinwedel and J. H. D. Jensen, Z. Naturforsch. 5a, 413 (1950).
[14] D. H. Wilkinson, Physica 22, 1039 (1956).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

In the internal region $r<R$, ϕ can be obtained as

$$\phi(r)=\sum a_m\phi_m(r), \quad r<R, \tag{3}$$

where the ϕ_m are eigenfunctions satisfying the equation

$$\frac{h^2}{2M}\frac{d^2\phi_m}{dr^2}+(E_m-V)\phi_m=0, \tag{3.1}$$

and the boundary condition

$$\left\{\frac{d\phi_m}{dr}=ik\phi_m\right\}\Bigg|_{r=R}. \tag{3.2}$$

The eigenvalues E_m are complex because of this imaginary boundary condition, the boundary condition depending explicitly on the wave number and hence, the energy of the incident neutron. We discuss the consequences of this later. These eigenfunctions form a complete set.[15] Orthogonality of the ϕ_m is easily established by considering the equation for ϕ_n,

$$\frac{h^2}{2M}\frac{d^2\phi_n}{dr^2}+(E_n-V)\phi_n=0, \tag{3.3}$$

multiplying Eq. (3.1) on the left by ϕ_n and Eq. (3.3) on the left by ϕ_m and subtracting. Thus,

$$\frac{h^2}{2M}\left(\phi_n\frac{d^2\phi_m}{dr^2}-\phi_m\frac{d^2\phi_n}{dr^2}\right)+(E_m-E_n)\phi_n\phi_m=0. \tag{3.4}$$

By integrating this equation from 0 to R and using Green's theorem, one finds that

$$\frac{h^2}{2M}\left(\phi_n\frac{d\phi_m}{dr}-\phi_m\frac{d\phi_n}{dr}\right)_{r=R}=(E_n-E_m)\int_0^R\phi_n\phi_m dr. \tag{3.5}$$

From Eq. (3.2) it can be seen that the left-hand side vanishes and hence

$$\int_0^R\phi_n(r)\phi_m(r)dr=0, \quad E_m\neq E_n. \tag{3.6}$$

In general, $E_m\neq E_n$ implies $m\neq n$; the exceptional case $E_n=E_m$ for $m\neq n$ can be handled by special methods, but it does not occur for the type of potentials we consider. For $m=n$, we choose the normalization of the ϕ_m's so that the integral is equal to unity, i.e.,

$$\int_0^R\phi_m^2(r)dr=1. \tag{3.7}$$

Two important features are first, the orthogonality is between ϕ_n and ϕ_m, not ϕ_n^* and ϕ_m as is usually the case, and secondly, the orthogonality depends essentially on the fact that ϕ_m and ϕ_n obey the same boundary condition.

[15] R. E. Peierls, Proc. Cambridge Phil. Soc. 44, 242 (1948).

By using the orthogonality of the ϕ_m we can obtain the a_m of Eq. (3) by again using Green's theorem.

$$\int_0^R(\phi_m\mathfrak{L}\phi-\phi\mathfrak{L}\phi_m)dr=-\frac{h^2}{2M}\left(\phi_m\frac{d\phi}{dr}-\phi\frac{d\phi_m}{dr}\right)_R, \tag{4}$$

with

$$\mathfrak{L}=-\frac{h^2}{2M}\frac{d^2}{dr^2}+V(r). \tag{4.1}$$

From Eqs. (1) and (3.1), one obtains

$$(E-E_m)a_m=-\frac{h^2}{2M}\left[\left(\frac{d\phi}{dr}-ik\phi\right)\phi_m\right]_R$$

$$=-\frac{h^2}{2M}e^{-ikR}\phi_m(R), \tag{4.2}$$

giving

$$a_m=\frac{h^2}{2M}\frac{e^{-ikR}}{E_m-E}\phi_m(R). \tag{4.3}$$

The joining of the inside and outside solutions at $r=R$ gives

$$\sum a_m\phi_m(R)=\frac{\sin kR}{k}+Se^{ikR}, \tag{5}$$

giving

$$S=\frac{1}{ke^{ikR}}\left\{\sum_m\frac{kh^2}{2M}e^{-ikR}\frac{[\phi_m(R)]^2}{E_m-E}-\sin kR\right\}. \tag{6}$$

This can be put into familiar form by defining the width

$$\Gamma_m\equiv\frac{kh^2}{M}[\phi_m(R)]^2. \tag{6.1}$$

Then

$$S=\frac{e^{-2ikR}}{2k}\sum_m\frac{\Gamma_m}{E_m-E}-\frac{\sin kR}{k}e^{-ikR}, \tag{7}$$

and

$$\sigma=4\pi|S|^2. \tag{7.1}$$

Because of the imaginary boundary condition, the ϕ_m and hence the Γ_m are complex. For low bombarding energies, the imaginary part is small as appears later.

The imaginary part of E_m can be found by the development

$$\int_0^R(\phi_m\mathfrak{L}\phi_m^*-\phi_m^*\mathfrak{L}\phi_m)dr$$

$$=(E_m^*-E_m)\int_0^R\phi_m\phi_m^*dr. \tag{8}$$

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

From Green's theorem, the left-hand side is

$$(ik\hbar^2/M)\phi_m^*\phi_m(R),$$

and if the imaginary part of E_m is denoted by $-\beta_m/2$, i.e.,

$$E_m=\epsilon_m-i\beta_m/2, \tag{8.1}$$

then from Eq. (8) we find that

$$\beta_m=\frac{k\hbar^2}{M}\frac{|\phi_m(R)|^2}{\int_0^R|\phi_m(r)|^2dr}. \tag{8.2}$$

This has a simple interpretation. The numerator is proportional to the escape velocity of the particle $k\hbar/M$ multiplied by the probability of the particle being at the surface, whereas the denominator represents the probability of the particle being in the nucleus. At low energies where the ϕ_m are mainly real, as already indicated, the denominator is nearly unity [see Eq. (3.7)] and comparing Eqs. (8.2) and (6.1), we see that

$$\beta_m\cong\Gamma_m \quad \text{(low energies)}. \tag{8.3}$$

To obtain a better idea of the sizes of the various quantities, it is useful to illustrate the formalism by evaluating them for a square well. We choose a well $V=-U$, $r<R_0$; $V=0$, $r>R_0$; with $U=42$ Mev and $R_0=1.45$ A$^{\frac{1}{3}}$ $\times10^{-13}$ cm, where A is the atomic number; these are the parameters of the well employed by Feshbach, Porter, and Weisskopf,[9] aside from the imaginary part employed by them which is introduced later. Consequently, these parameters are typical of those considered later. The solution of Eqs. (3.1) and (3.2) are compared later with $\phi_m^{(0)}$ defined by

$$\frac{\hbar^2}{2M}\frac{d^2}{dr^2}\phi_m^{(0)}(r)+(E_m^{(0)}-V)\phi_m^{(0)}(r)=0, \tag{9}$$

$$\left.\frac{d\phi_m^{(0)}}{dr}\right|_{r=R}=0. \tag{9.1}$$

This real boundary condition is that of the Wigner-Eisenbud theory, and the solutions $\phi_m^{(0)}$ are real. The functions $\phi_m^{(0)}$ are considered because it is easier to calculate these real functions and, at low energies, they are a good zero-order approximation to ϕ_m as we show.

In the square well ϕ_m is equal to

$$\phi_m=A_m\sin(K_m-i\kappa_m)r, \tag{9.2}$$

with K_m and κ_m defined by

$$(K_m-i\kappa_m)^2=2M(E_m+U)/\hbar^2. \tag{9.3}$$

For simplicity, the joining radius R in Eq. (3.2) has been chosen to be the edge of the well, R_0. It can be chosen differently and is generally chosen somewhat larger, but this is immaterial for the qualitative features of the results derived here. The boundary condition, Eq. (3.2), can now be written

$$(K_m-i\kappa_m)\cos(K_m-i\kappa_m)R=ik\sin(K_m-i\kappa_m)R. \tag{9.4}$$

Now κ_mR is shown to be small at low energies; Eq. (9.4) is expanded to first order in this quantity. Then

$$(K_m-i\kappa_m)(\cos K_mR+i\kappa_mR\sin K_mR)\cong ik(\sin K_mR-i\kappa_mR\cos K_mR). \tag{9.5}$$

This gives

$$\kappa_m=k\sin K_mR/(K_mR\sin K_mR-\cos K_mR). \tag{9.6}$$

Since $\cos K_mR\ll K_mR\sin K_mR$ [see Eq. (9.8)], it follows that

$$\kappa_m\cong k/K_mR. \tag{9.7}$$

By using this value for κ_m, one obtains from Eq. (9.5)

$$\cos K_mR\cong-(kR)^2\sin K_mR/(K_mR)^3 \tag{9.8}$$

to lowest order in κ_mR. Now K_mR is large, of the order of 10 even for medium A and zero incident energy, and so κ_m and $\cos K_mR$ are small. Equation (9.7) shows that κ_mR is small—which is necessary for the rapid convergence of the expansion above—as long as $E\ll U$, that is, as long as the bombarding energy is much less than the depth of the well.

We now let $K_mR=K_m^{(0)}R+\delta K_mR$, where $\phi_m^{(0)}=(2/R)^{\frac{1}{2}}\sin K_m^{(0)}R$, and $K_m^{(0)}R=(n+\frac{1}{2})\pi$ gives a solution to Eqs. (9) and (9.1). To lowest order in κ_mR, Eq. (9.8) shows that

$$\delta K_m=k[kR/(K_mR)^3]. \tag{9.9}$$

Although this is a second-order correction (in the sense that δK_mR is of second order in kR), it is given correctly to this order by our first-order expression Eq. (9.5), since further terms enter only in third order. We find, further, that

$$\delta\epsilon_m\equiv\epsilon_m-E_m^{(0)}=(\hbar^2/2M)(2K_m\delta K_m-\kappa_m^2)\cong[1/(K_mR)^2]\epsilon_m, \tag{9.10}$$

where the energies are measured from the bottom of the well. For a typical value of K_mR of ~10, this means that calculation of the real part of the resonance energy from the boundary condition Eq. (9.1) gives results accurate to $\sim1\%$.

The imaginary part β_m of E_m is given to first order in κ_mR by $(\hbar^2k/M)[\phi_m^{(0)}(R)]^2$ which is equal to

$$\beta_m\cong2\hbar^2k/MR\cong\Gamma_m, \tag{9.11}$$

since

$$\phi_m^{(0)}(r)=(2/R)^{\frac{1}{2}}\sin K_mr. \tag{9.12}$$

Thus, to lowest order in κ_mR, the Γ_m are real and the dependence of the real parts of the resonance energies ϵ_m on the bombarding energy enters only in the second order of this quantity.

Generalization of these results to the case of a

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

complex square well is trivial. The introduction of $\tilde{V}(r)=-U-iW$ (use $\tilde{V}$ to denote a complex well; eigenfunctions in this complex well are denoted by $\tilde{\phi}_m(r)$, etc.) so that W simply shifts the resonance energy by the constant amount $-iW$. Hence,

$$\tilde{E}_m=E_m-iW=\epsilon_m-i\beta_m/2-iW. \quad (9.13)$$

Consequently, introduction of an imaginary part broadens the resonance levels. The eigenfunctions are unchanged, i.e., $\tilde{\phi}_m(r)=\phi_m$ in the special case of the square well. The development for the case in which the real and imaginary parts vary with r is easily carried out. This development can also be carried out for complex velocity-dependent potentials[16]; this is important since the parameters of the optical model wells occurring in practice depend on the bombarding energy. It is convenient to take the depth of the well to be different for the different eigenenergies at a given bombarding energy. In discussing a velocity-dependent well $\tilde{V}(r)$, one can consider $\tilde{V}$ to be a nonlocal operator $\tilde{V}(\mathbf{r}',\mathbf{r})$, which is equivalent to a velocity-dependent potential. The wave equation is

$$\frac{\hbar^2}{2M}\frac{d^2}{dr'^2}\tilde{\phi}_m(r')+\tilde{E}_m\tilde{\phi}_m(r')-\int\tilde{V}(\mathbf{r}',\mathbf{r})\tilde{\phi}_m(r)dr=0. \quad (10)$$

Orthogonality of the $\tilde{\phi}_m$ requires $\tilde{V}(\mathbf{r}',\mathbf{r})=\tilde{V}(\mathbf{r},\mathbf{r}')$; potentials not obeying this relation can be shown to be physically unreasonable.[16] Proof of orthogonality assumes $\tilde{V}(\mathbf{r},\mathbf{r}')$ to be zero unless both r and r' are less than R.

The resonance treatment for a neutron in the complex potential has been developed in some detail, not only to illustrate the formalism, but also because knowledge of the positions, widths, and spacings of the levels is often useful in theoretical estimates.

Generalization of these results to the case of nonzero angular momentum is easy. In this case we express the function, regular at the origin, $\psi_\alpha(\mathbf{r})$ as

$$r\psi_\alpha(\mathbf{r})=\Theta_\alpha(\theta,\varphi)\phi_\alpha(r), \quad (11)$$

where $\Theta_\alpha(\theta,\varphi)$ is the normalized function of angles, e.g., in the case of spinless particles, $\Theta_\alpha(\theta,\varphi)$ is equal to $Y_l^m(\theta,\varphi)$, where Y_l^m is the normalized spherical harmonic, and ϕ_α is the radial function in the potential well, regular at the origin and asymptotic to

$$e^{i\delta_l}\sin[kr-(l\pi/2)+\delta_l],$$

where δ_l is the phase shift for the lth wave. The lower index α labels all angular momentum quantum numbers, i.e., in the terminology of Wigner and co-workers, it is the "channel" index. Further, solutions asymptotic to outgoing waves are denoted by $\psi^+(\mathbf{r})$, i.e.,

$$r\psi_\alpha^+(\mathbf{r})=\Theta_\alpha(\theta,\varphi)\phi_\alpha^+(r), \quad (11.1)$$

where $\phi_\alpha^+(r)$ is asymptotic to $\exp(ikr-il\pi/2)$ in the case of neutrons. Generalization to the case of charged particles is easily made. Here $\phi_\alpha^+(r)$ is asymptotic to $\exp(ikr-\eta\ln 2kr-l\pi/2+\sigma_l)$ where $\eta=Ze^2/\hbar v$, with v the velocity of the proton, and σ_l is the Coulomb phase shift. Solution of the radial Schrödinger equation can be expressed for $r>R$, as

$$\phi_\alpha(r)=[\phi_\alpha^+(r)-\phi_\alpha^-(r)]/2ik+S_\alpha\phi_\alpha^+(r), \quad (12)$$

with $\phi_\alpha^-=(\phi_\alpha^+)^*$, the solution asymptotic to incoming waves. The eigenfunctions ϕ_m (these have the same angular dependence as the $\phi_\alpha(r)$, so we do not carry the subscript α on them) are then determined by the boundary condition

$$\left[\frac{d\phi_m}{dr}=f_\alpha^+(r)\phi_m\right]_{r=R} \quad (12.1)$$

where

$$f_\alpha^+(r)=\frac{1}{\phi_\alpha^+(r)}\frac{d\phi_\alpha^+(r)}{dr} \quad (12.2)$$

is the logarithmic derivative for an outgoing wave at the joining radius. Again, $\phi_\alpha(r)$ can be expanded in terms of the ϕ_m,

$$\phi_\alpha(r)=\sum_m a_m\phi_m(r), \quad r<R, \quad (12.3)$$

and the development analogous to that of Eqs. (4) to (4.3) now gives

$$(E-E_m)a_m=-\frac{\hbar^2}{2M}\left[\left(\frac{d\phi_\alpha}{dr}-f^+\phi_\alpha\right)\phi_m\right]_R. \quad (12.4)$$

The use of Eq. (12) for ϕ_α leaves only the term in ϕ_α^- on the right-hand side of Eq. (12.4), and we obtain

$$(E-E_m)a_m=\frac{\hbar^2}{4Mik}[(f_\alpha^--f_\alpha^+)\phi_\alpha^-\phi_m]_R. \quad (12.5)$$

The Wronskian is

$$[f_\alpha^+(r)-f_\alpha^-(r)]\phi_\alpha^+(r)\phi_\alpha^-(r)=2ik, \quad (12.6)$$

so that we obtain

$$a_m=\frac{\hbar^2}{2M}\frac{1}{\phi_\alpha^+(R)}\frac{\phi_m(R)}{(E_m-E)}. \quad (12.7)$$

The joining of the inside and outside solutions at $r=R$ now gives

$$S_\alpha=\frac{1}{k[\phi_\alpha^+(R)]^2}\sum_m\frac{k\hbar^2}{2M}\frac{\phi_m^2(R)}{E_m-E}-\frac{\phi_\alpha^+(R)-\phi_\alpha^-(R)}{2ik\phi_\alpha^+(R)}. \quad (12.8)$$

We define

$$\Gamma_m\equiv\frac{k\hbar^2}{M}P(R)\phi_m^2(R), \quad (12.9)$$

[16] See Appendix C of G. E. Brown and C. T. De Dominicis, Proc. Phys. Soc. (London) A72, 70 (1958).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

where P is the penetrability,

$$P=\frac{1}{\phi_\alpha^+(R)\phi_\alpha^-(R)}=\frac{1}{|\phi_\alpha^+(R)|^2}. \tag{12.10}$$

We then have

$$S_\alpha=\frac{1}{2k}\frac{\phi_\alpha^-(R)}{\phi_\alpha^+(R)}\sum_m\frac{\Gamma_m}{E_m-E}-\frac{1}{2ik}\left(1-\frac{\phi_\alpha^-(R)}{\phi_\alpha^+(R)}\right). \tag{12.11}$$

The cross section is given by

$$\sigma_\alpha=4\pi|S_\alpha|^2$$
$$=\frac{\pi}{k^2}\left|\frac{\phi_\alpha^-(R)}{\phi_\alpha^+(R)}\sum_m\frac{\Gamma_m}{E_m-E}+i\left(1-\frac{\phi_\alpha^-(R)}{\phi_\alpha^+(R)}\right)\right|^2. \tag{12.12}$$

2. The Many-Body Case

We consider the case in which an incident nucleon is scattered either elastically or inelastically by a nucleus of A particles.[17] Once this particle is inside the radius R —which is chosen so that the interaction between the incident particle and the nucleus vanishes outside this radius—the wave function can again be expanded in compound states $\Phi^{(p)}(\mathbf{r},\xi)$. Here we use $\mathbf{r}$ for the coordinate of the incident nucleon and ξ to label the totality of coordinates of the A particles inside the nucleus. For the moment the incident particle is treated as distinguishable from those in the nucleus, but later the generalization to the case where it is identical with particles in the nucleus is indicated.

The compound states $\Phi^{(p)}(\mathbf{r},\xi)$ obey the equations

$$H\Phi^{(p)}(\mathbf{r},\xi)=W_p\Phi^{(p)}(\mathbf{r},\xi),$$
$$H=H_\xi+T(r)+V(\mathbf{r},\xi). \tag{13}$$

H_ξ is the Hamiltonian of the A nuclear particles, $T(r)$ is the kinetic energy of the incident particle, and $V(\mathbf{r},\xi)$ is the potential interaction between the A nuclear particles and the incident nucleon. It is a sum of nucleon-nucleon potentials,

$$V(\mathbf{r},\xi)=\sum_{i=1}^{A}V(\mathbf{r},\mathbf{r}_i),$$

the nucleon to which the coordinate $\mathbf{r}$ refers being the incident particle. It is assumed that $V(\mathbf{r},\xi)$ is a well-behaved potential, i.e., a potential without a strong repulsive core and other singularities; generalization to the case that it is not can be made using techniques developed by Watson and Brueckner, but this only introduces nonessential complications for the points considered here.

[17] The emission of composite particles is dropped from the dispersion-theory description later on when one makes the assumption that the interaction between the incident particle and all of the A particles vanishes beyond the joining radius R. These processes can presumably be included by modifying the formalism.

States $\Phi^{(p)}$ and $\Phi^{(q)}$ are not now orthogonal, but rather $\Phi^{(p)}$ and $\bar{\Phi}^{(q)}$, where $\bar{\Phi}^{(q)}$ is obtained from $\Phi^{(q)}$ by taking the complex conjugate of all functions of angles, in particular, of the $\Theta_\alpha(\theta,\varphi)$ occurring in the single-particle function. Equivalently, to within an arbitrary phase factor, $\bar{\Phi}^{(q)}$ is obtained from $\Phi^{(q)}$ by rotating the wave function so that the total angular momentum component M is changed into $-M$ without taking the complex conjugate of the function of intrinsic structure (e.g., the radial wave function in the single-particle case). We choose

$$\int_0^R\bar{\Phi}^{(q)}\Phi^{(p)}d^3\xi d^3r=1,\quad(p=q), \tag{13.1}$$

and the integral is zero for $p\neq q$. In using the orthogonality later on, we do not indicate the bar over the Φ on the left since the change in the angular momentum functions is a trivial one to introduce once we introduce the expansion, Eq. (14), and all integrals over angles are easy to carry out.

We introduce a complete set of states $\chi_j(\xi)$ for the A particles. Boundary conditions for these states could be chosen in various ways. However, the tightly bound states close to the ground state in which we are mainly interested are insensitive to the precise boundary condition, e.g., if we choose the joining radius R fairly far out, then both the wave function and its derivative are small for a bound state defined in any sensible way, and it would not matter much whether we chose the former or the latter to be zero, or to have some small finite value at that radius as our boundary condition. Difficulties enter when the excitation energy ϵ_j, which we measure from ϵ_0 as origin, of the state χ_j is sufficiently large so that one of the A particles can escape from the nucleus. However, since the only property of these highly excited states that we use is that they, together with those of low excitation, form a complete set, we believe our results to be independent of the way in which these boundary conditions are chosen.

In order to make the connection with the optical model later, it is convenient to employ states $\tilde{\psi}_m(\mathbf{r})$ in a complex well of the type discussed in the last section for the incident particle. We can then expand

$$\Phi^{(p)}(\mathbf{r},\xi)=\sum_{j,m}a_{jm}{}^p\chi_j(\xi)\tilde{\psi}_m{}^j(\mathbf{r}). \tag{14}$$

The boundary condition on $\Phi^{(p)}$ at $r=R$ can now be simply stated in terms of the boundary conditions on the $\tilde{\psi}_m{}^j$. It is convenient, in analogy with Eq. (11), to introduce radial functions $\tilde{\phi}_m$ through

$$r\tilde{\psi}_m{}^j(\mathbf{r})=\Theta_\alpha(\theta,\varphi)\tilde{\phi}_m{}^j(r). \tag{14.1}$$

The boundary condition is then

$$\left[\frac{d\tilde{\phi}_m{}^j}{dr}=f_\alpha^+(r,E-\epsilon_j)\tilde{\phi}_m{}^j(r)\right]_{r=R}. \tag{14.2}$$

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

The $E-\epsilon_j$ in the argument of f_α^+ indicates that it is to be taken for that energy. The boundary condition is chosen to be that of an outgoing wave at the reduced energy available to the outgoing particle after leaving the residual nucleus in excited state j. The channel index α on the $\tilde{\psi}_m{}^j$ is again suppressed. Because of the dependence of f_α^+ on ϵ_j, the $\tilde{\psi}$'s also depend on j, as indicated; however, this dependence is weak. The fact that the boundary condition on the $\tilde{\psi}$'s depends on j does not upset the orthogonality of the $\Phi^{(p)}$'s as long as one always integrates over the ξ coordinates first.

We can express the wave function $\Psi(\mathbf{r},\xi)$, which is the solution to the Schrödinger equation,

$$H\Psi=E\Psi, \tag{15}$$

as

$$\Psi(\mathbf{r},\xi)=\sum a_p\Phi^{(p)}(\mathbf{r},\xi) \quad \text{for } r<R, \tag{15.1}$$

and as

$$\Psi(\mathbf{r},\xi)=I_\alpha\left(\frac{\psi_\alpha^+(\mathbf{r})-\psi_\alpha^-(\mathbf{r})}{2ik}\right)\chi_o(\xi)+S_\alpha\psi_\alpha^+(\mathbf{r})\chi_o(\xi)$$
$$+\sum_{\alpha\neq\alpha'} S_{\alpha\alpha',j}\psi_{\alpha'}{}^{j+}(\mathbf{r})\chi_j(\xi) \quad \text{for } r>R, \tag{15.2}$$

where S_α is the amplitude for elastic scattering in channel α and $S_{\alpha\alpha',j}$ is the amplitude for inelastic scattering from channel α into channel α', leaving the nucleus excited in state χ_j. The $\psi_{\alpha'}{}^{j+}$ represents an outgoing wave at energy $E-\epsilon_j$ in channel α'. The coefficient I_α represents the amplitude for channel α in the plane wave. In the case of spinless particles,

$$e^{i\mathbf{k}\cdot\mathbf{r}}=\sum_{l=0}^{\infty} i^l[4\pi(2l+1)]^{\frac{1}{2}}Y_l{}^0(\theta,\varphi)j_l(kr). \tag{15.3}$$

If we set $\alpha=l$, i.e., identify $Y_l{}^0$ with Θ_α for this case, and use

$$j_l\sim\frac{\sin(kr-l\pi/2)}{kr}, \tag{15.4}$$

we choose

$$I_l=i^l(2l+1)^{\frac{1}{2}} \tag{15.5}$$

for this special case. The $(4\pi)^{\frac{1}{2}}$ is taken into account at a later stage by multiplying by 4π to obtain the cross section, as in Eqs. (7.1) and (12.12); this convention makes our general treatment correspond to our S-wave case when $l=0$.

We can again find the a_p by Green's theorem,

$$\int_0^R(\Phi^{(p)}H\Psi-\Psi H\Phi^{(p)})d^3\xi d^3r=(E-W_p)a_p$$
$$=-\frac{h^2}{2M}\sum_{j,m}a_{jm}{}^p\int_0^R d^3\xi\chi_j(\xi)$$
$$\times\int d\Omega\left\{R\tilde{\psi}_m{}^j(\mathbf{R})\frac{d}{dr}[r\Psi(\mathbf{r},\xi)]\right.$$
$$\left.-R\Psi(\mathbf{R},\xi)\frac{d}{dr}[r\tilde{\psi}_m{}^j(\mathbf{r})]\right\}_{r=R}. \tag{16}$$

With the use of Eq. (14.2), we find

$$a_p=\frac{h^2}{2M(W_p-E)}\sum_{j,m}a_{jm}{}^p\int_0^R d^3\xi\chi_j(\xi)\int d\Omega$$
$$\times\left\{R\tilde{\psi}_m{}^j(\mathbf{R})\left[\frac{d}{dr}-f_\alpha^+(R,E-\epsilon_j)\right]r\Psi(\mathbf{r},\xi)\right\}_{r=R}, \tag{16.1}$$

where we have indicated the energy dependence of the f_α^+ explicitly. The operator in [ ]'s in Eq. (16.1) gives zero operating on any "outgoing" part of Ψ, and the only term that contributes in Eq. (15.2) is the term in $\psi_\alpha^-(\mathbf{r})\chi_o(\xi)$. Hence, using the orthogonality of the χ's,

$$a_p=-I_\alpha\frac{h^2}{2M(W_p-E)}\sum_m a_{om}{}^p\left\{\frac{f_\alpha^-(R)-f_\alpha^+(R)}{2ik}\right\}$$
$$\times\int d\Omega R\tilde{\psi}_m(\mathbf{R})R\psi_\alpha^-(\mathbf{R}). \tag{16.2}$$

The integral over Ω guarantees that the angular dependence of the $\tilde{\psi}_m$ is the same as that of the incident channel α. With this restriction understood, we can write a_p in terms of the radial part of the wave functions,

$$a_p=I_\alpha\frac{h^2}{2M(W_p-E)}\left\{\frac{f_\alpha^--f_\alpha^+}{2ik}\right\}\phi_\alpha^-(R)\sum_m a_{om}{}^p\tilde{\phi}_m(R). \tag{16.3}$$

We can simplify Eq. (16.3) by using the Wronskian, Eq. (12.6), giving

$$a_p=I_\alpha\frac{h^2}{2M(W_p-E)}\frac{1}{\phi_\alpha^+(R)}\sum_m a_{om}{}^p\tilde{\phi}_m(R). \tag{16.4}$$

By having a_p, we can now obtain S_α and $S_{\alpha\alpha',j}$ by matching the internal and external wave functions, Eqs. (15.1) and (15.2), at the joining radius. If both equations are multiplied by $\chi_o(\xi)$ and $R\Theta_\alpha$ and integrated over $d^3\xi$ and $d\Omega$, one obtains

$$S_\alpha\phi_\alpha^+(R)=\sum_p a_p\sum_{m'}a_{om'}{}^p\tilde{\phi}_{m'}(R)$$
$$-I_\alpha(\phi_\alpha^+(R)-\phi_\alpha^-(R))/2ik. \tag{17}$$

If we introduce the width,

$$\gamma_p=\frac{kh^2}{M}\frac{1}{\phi_\alpha^+(R)\phi_\alpha^-(R)}\sum_m\sum_{m'}a_{om}{}^pa_{om'}{}^p\tilde{\phi}_m(R)\tilde{\phi}_{m'}(R), \tag{17.1}$$

we can write

$$S_\alpha=\frac{I_\alpha\phi_\alpha^-(R)}{2k\,\phi_\alpha^+(R)}\sum_p\frac{\gamma_p}{W_p-E}-\frac{I_\alpha}{2ik}\left(1-\frac{\phi_\alpha^-(R)}{\phi_\alpha^+(R)}\right), \tag{18}$$

which is of the same form as Eq. (12.11). The cross

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

section σ_α for elastic scattering is given by

$$\sigma_\alpha = 4\pi |S_\alpha|^2. \tag{18.1}$$

For spinless particles, α can be replaced by the label l, and with the value of I_l given by Eq. (15.5), one obtains

$$\sigma_l = (2l+1)\frac{\pi}{k^2}\left|\frac{\phi_l^-(R)}{\phi_l^+(R)}\sum_p \frac{\gamma_p}{W_p - E} + i\left(1 - \frac{\phi_l^-(R)}{\phi_l^+(R)}\right)\right|^2, \tag{18.2}$$

for the total scattering through channel l. Since the incident particle is described by a plane wave, Eq. (15.3), the differential cross section is given as the square of a sum of terms involving different l,[18]

$$\frac{d\sigma(\theta)}{d\Omega} = 4\pi \left|\sum_l i^{-l} S_l Y_l^0\right|^2. \tag{18.3}$$

Similarly, the amplitude for inelastic scattering is

$$S_{\alpha\alpha',j} = \frac{1}{\phi_{\alpha'}{}^{j+}(R)} \sum_p a_p \sum_m a_{jm}{}^p \tilde{\phi}_m{}^j(R). \tag{19}$$

In the inelastic case, the widths do not occur naturally in the amplitude, but if quantities u_p and u_{pj} are defined by

$$u_p = \left(\frac{kh^2}{M}\right)^{\frac{1}{2}} \frac{1}{|\phi_\alpha{}^+(R)|} \sum_m a_{om}{}^p \tilde{\phi}_m(R), \tag{19.1}$$

and

$$u_{pj} = \left(\frac{k_j h^2}{M}\right)^{\frac{1}{2}} \frac{1}{|\phi_{\alpha'}{}^{j+}(R)|} \sum_m a_{jm}{}^p \tilde{\phi}_m{}^j(R), \tag{19.1}$$

then

$$\sigma_{\alpha\alpha',j} = \frac{4\pi k_j}{k}|S_{\alpha\alpha',j}|^2 = \frac{\pi}{k^2}\left|I_\alpha \sum_p \frac{u_p u_{pj}}{W_p - E}\right|^2. \tag{19.2}$$

The ratio k_j/k, with k_j defined by $k_j{}^2 = 2M(E-\epsilon_j)/\hbar^2$, is the ratio of the velocity of the outgoing particle to that of the incident one, which enters into the cross section.

In case one term, p, in the sum in Eq. (19.2) contributes the main part of the sum, the inelastic cross section can be put into the form of the Breit-Wigner formula

$$\sigma_{\alpha\alpha',j} = \frac{\pi}{k^2}|I_\alpha|^2 \frac{\gamma_p \gamma_{pj}}{|W_p - E|^2}. \tag{19.3}$$

(See the Breit-Wigner formula given in the introduction; there we indicated the channel index on the γ's explicitly.) In Eq. (19.3)

$$\gamma_p = u_p{}^2,$$

as previously, and

$$\gamma_{pj} = u_{pj}{}^2 \tag{19.4}$$

is the width for the inelastic process.

Now $W_p = \epsilon_p - i\alpha_p/2$, where ϵ_p and α_p are real, and we are interested in the width α_p. This can be obtained from the original and conjugate equations,

$$(H - W_p)\Phi^{(p)} = 0, \tag{20}$$
$$(H - W_p{}^*)\Phi^{(p)*} = 0.$$

If the first equation is multiplied on the left by $\Phi^{(p)*}$ and the second by $\Phi^{(p)}$, one obtains, after subtracting and integrating,

$$i\alpha_p \int_0^R \Phi^{(p)*}\Phi^{(p)} d^3\xi d^3r = \frac{\hbar^2}{2M}\int d^3\xi \int d\Omega \times \left[r\Phi^{(p)*}\frac{d}{dr}(r\Phi^{(p)}) - r\Phi^{(p)}\frac{d}{dr}(r\Phi^{(p)*})\right]_{r=R}. \tag{21}$$

By using the expansion Eq. (14) and the boundary conditions Eq. (14.2), this becomes

$$i\alpha_p \int_0^R \Phi^{(p)*}\Phi^{(p)} d^3\xi d^3r = \frac{\hbar^2}{2M}\sum_{j,m,m'} a_{jm}{}^{p*} a_{jm'}{}^p \tilde{\phi}_m{}^{j*}(R)\tilde{\phi}_m{}^j(R) \times [f_{\alpha'}{}^+(E-\epsilon_j, R) - f_{\alpha'}{}^-(E-\epsilon_j, R)]. \tag{22}$$

Use of the Wronskian, Eq. (12.6), results in

$$\alpha_p = \frac{\hbar^2}{M}\sum_{j,m,m'} \frac{k_j a_{jm}{}^{p*} a_{jm'}{}^p \tilde{\phi}_m{}^{j*}(R)\tilde{\phi}_{m'}{}^j(R)/\phi_{\alpha'}{}^{j+}(R)\phi_{\alpha'}{}^{j-}(R)}{\int_0^R \Phi^{(p)*}\Phi^{(p)} d^3\xi d^3r}, \tag{22.1}$$

where the sum over j includes the term $j=0$. The right-hand side of Eq. (22.1) has a simple interpretation similar to that of Eq. (8.2). In the numerator the jth term in the sum is proportional to the probability of the particle being at the radius R, multiplied by its escape velocity and the relevant penetrability. The integral in the denominator represents the probability of the particle in compound state p being in the nucleus. (The normalization is such that this probability is not

[18] The factor i^{-l} enters because S_l is the coefficient of $\exp(ikr - il\pi/2)$, so that the coefficient of $\exp ikr$ is

$$S_l \exp(-il\pi/2) = S_l i^{-l}.$$

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

necessarily unity.) In the low-energy region where all of the quantities on the right-hand side of Eq. (22.1) are essentially real, as is shown later, this reduces to

$$\alpha_p=\sum_j \gamma_{pj}. \tag{22.2}$$

This is interpreted as the total width being equal to the sum of partial widths for the various processes.

Equation (18) is similar to a many-level Breit-Wigner formula. This relationship between the scattering amplitude and the sum of resonance terms is linear, and this is why it is often simpler to employ in the many-level case than the Wigner-Eisenbud expressions. However, before it can be applied to the case in which only one or two levels contribute, as in the application to low-energy reactions with light nuclei, it is sometimes necessary to make the energy dependence of the γ_p's (which results from the dependence of the boundary conditions on E) explicit, and to determine the imaginary part [the γ_p's defined by Eq. (17.1) are complex]. The determination of these is equivalent to relating the formalism back to the Wigner-Eisenbud one in which the widths are real and independent of energy. The approximate relationship was developed in perturbation theory by Kapur and Peierls,[3] but we prefer to use a procedure developed by A. M. Lane.[19] States $\tilde{\psi}_m{}^{j(0)}(\mathbf{r})$ are defined for which the radial part, multiplied by r, obeys the energy-independent boundary condition

$$\left[\frac{d}{dr}r\tilde{\phi}_m{}^{j(0)}(r)=0\right]_{r=R}, \tag{23}$$

in analogy with Eq. (9.1). By building up $\Phi_0{}^{(q)}$ as

$$\Phi_0{}^{(q)}(\mathbf{r},\xi)=\sum_{j,m} a_{jm}{}^{q(0)}\chi_j(\xi)\tilde{\psi}_m{}^{j(0)}(\mathbf{r}), \tag{23.1}$$

we define energy-independent compound states. The overlap between such states and our energy-dependent compound states can be found by employing Green's theorem again as in Eq. (16). This gives

$$(W_p-W_q{}^{(0)})\int_0^R \Phi_0{}^{(q)}\Phi^{(p)}d^3\xi d^3r$$
$$=-\frac{h^2}{2M}\int_0^R d^3\xi\int d\Omega\left\{R\Phi_0{}^{(q)}(\mathbf{R},\xi)\times\frac{d}{dr}[r\Phi^{(p)}(\mathbf{r},\xi)]\right\}_{r=R}. \tag{23.2}$$

For the case in which the resonances p are well separated and only S-wave scattering can occur, we have

$$(W_p-W_q{}^{(0)})\int_0^R \Phi_0{}^{(q)}\Phi^{(p)}d^3\xi d^3r$$
$$=-\frac{ih^2}{2M}\sum_{j,m,n} k_j a_{jm}{}^{q(0)}a_{jn}{}^{p}\tilde{\phi}_m{}^{j(0)}(R)\tilde{\phi}_n{}^{j}(R). \tag{24}$$

This equation is exact, but we now use perturbation theory starting with the zero-order approximation $\Phi^{(p)}=\Phi_0{}^{(p)}$. To obtain the first-order correction to the energy, we let $p=q$ and set $\Phi_0{}^{(p)}$ equal to $\Phi^{(p)}$ under the integral. This gives

$$W_p-W_p{}^{(0)}\cong -i\alpha_p/2,$$

where we have used Eq. (22.1) and remembered that the $a_{jn}{}^{(p)}$ and $\tilde{\phi}_m{}^j$ are real in this approximation. Hence, the first-order correction to the energy simply adds an imaginary part which is just the width of the state $\Phi^{(p)}$ and which is independent of energy in this order. We can obtain the first-order correction to the wave function by expanding

$$\Phi^{(p)}=\Phi_0{}^{(p)}+\sum c_q{}^p\Phi_0{}^{(q)}, \tag{24.1}$$

and we obtain

$$(W_p-W_q{}^{(0)})c_q{}^p=-\frac{ih^2}{2M}\sum_{j,m,n} k_j a_{jm}{}^{q(0)}a_{jn}{}^{p}\times\tilde{\phi}_m{}^{j(0)}(R)\tilde{\phi}_n{}^{j}(R), \quad (p\neq q). \tag{24.2}$$

The right-hand side is of order $-i\alpha_p/2$, its approximate size for $p=q$, or less. (The signs of the $a_{jm}{}^{q(0)}$ and the $a_{jn}{}^{p}$ can be both positive and negative, and this tends to make the right-hand side smaller than $-i\alpha_p/2$. If a random-phase approximation were applicable, the right-hand side would be zero.) Therefore, for order of magnitude we have

$$c_q{}^p\sim -i\alpha_p/(W_p-W_q{}^{(0)}). \tag{24.3}$$

Thus in the case of well-separated levels, when

$$|W_p-W_q{}^{(0)}|\gg\alpha_p, \tag{24.4}$$

the $c_q{}^p$ and, consequently, the imaginary part of $\Phi^{(p)}$ are small, and the latter is of the order of the ratio of the width to the spacing of the levels. Hence, to a good approximation, the energy dependence of the various parameters and the imaginary parts of the widths are absent in the region of well-separated resonances, so that, except for effects especially sensitive to these features such as the "Thomas Shift," one can usually ignore them.

The arguments in the next section do not neglect the energy dependence or the imaginary parts of the widths; these are included correctly, although we may often use the case of well-separated levels, which is especially simple, as an illustration. These features do not essentially complicate the arguments.

III. THE OPTICAL MODEL AND DIRECT INTERACTION

1. The Physical Picture

The physical picture underlying the optical model was developed chiefly by Feshbach, Porter, and Weisskopf[9]

[19] A. M. Lane (private communication).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

and Friedman and Weisskopf.[20] These authors split the scattering amplitude S_α into

$$S_\alpha=\langle S_\alpha\rangle_{Av}+(S_\alpha-\langle S_\alpha\rangle_{Av}), \tag{25}$$

where $\langle S_\alpha\rangle_{Av}$ is the value of S_α averaged over an energy interval I. The averaging process is defined more precisely later. The average elastic cross section can now be expressed as

$$\langle\sigma_\alpha\rangle_{Av}=|\langle S_\alpha\rangle_{Av}|^2+\langle|S_\alpha-\langle S_\alpha\rangle_{Av}|^2\rangle_{Av}, \tag{26}$$

since the cross term between $\langle S_\alpha\rangle_{Av}$ and $S_\alpha-\langle S_\alpha\rangle_{Av}$ averages out. Although both terms on the right-hand side represent elastic scattering, they are physically quite different in nature.

Friedman and Weisskopf,[20] who used a time-dependent treatment, showed that if one constructs a wave packet of width ΔE in energy, then the scattered particles corresponding to the average amplitude $\langle S_\alpha\rangle_{Av}$, i.e., those given by the first term on the right-hand side of Eq. (26), pass over the nucleus in time $\tau=\hbar/\Delta E=\hbar/I$. This is, in fact, clearly allowed by the uncertainty principle, because one has, by averaging, constructed a scattering amplitude $\langle S_\alpha\rangle_{Av}$ which varies appreciably only over an energy interval I; since the energy of particles associated with this scattering is undefined within this interval, the time τ which the particle spends in the nucleus can be defined to within $\hbar/I$. Hence, if the energy interval I is very large, the particles corresponding to the average amplitude go over the nucleus very quickly. It is therefore reasonable that for sufficiently large intervals I, the particles make only a few interactions with the nuclear particles—even though the nucleon-nucleon potentials are relatively strong—and that we can describe them in the weak-coupling type picture indicated by the optical model. This scattering, which corresponds to the average phase, is termed "shape-elastic." Later we find precisely how large the interval I must be for various weak-coupling type descriptions of the average scattering to be valid.

On the other hand, the fluctuation scattering (often called compound-elastic scattering), described by the second term on the right-hand side of Eq. (26), varies over energy intervals of the order of the width of the compound nuclear states ($\sim$1 ev at low energies). Hence, the corresponding particles stay in the nucleus the order of a million times longer than those of the shape-elastic scattering that would correspond to a wave packet of the order of 1 Mev wide. If the particles corresponding to the shape-elastic scattering have time for one or two collisions, then those corresponding to the fluctuation scattering stay in the nucleus long enough to make the order of a million collisions. The latter particles can be regarded as those forming compound states of the type envisaged in the old compound-nucleus picture, whereas the shape-elastic scattering corresponds to that from the complex well, the imaginary part of which describes the disappearance of particles out of the incident beam into the long-lived component.

We define the complex well $\tilde{V}(r)$ so that it reproduces the average scattering phase, i.e., so that $\tilde{S}_\alpha=\langle S_\alpha\rangle_{Av}$. The physical picture indicates that this is the most reasonable procedure. We then go on to discuss the characteristics of this well $\tilde{V}(r)$ and relate the parameters back to nucleon-nucleon forces.

2. The Picture of Lane, Thomas, and Wigner

Possibly the most striking feature of experimental data on elastic neutron scattering was the appearance of the giant resonances in the total cross section in the energy region $\sim$0 to 3 Mev, as discussed in Sec. I. In order to understand these, we develop the picture of Lane, Thomas, and Wigner.[21] In this development we assume that the well $\tilde{V}(r)$, which reproduces the average scattering amplitude ($\tilde{S}_\alpha=\langle S_\alpha\rangle_{Av}$), has the characteristics of the optical model potential of Feshbach, Porter, and Weisskopf[9]; that is, if $\tilde{V}=-U-iW$ then $U\cong 42$ Mev, W lies in the range 1 to 2 Mev, and the radius of the potential is $R\cong 1.45\,A^{\frac{1}{3}}\times 10^{-13}$ cm. These parameters are used only for order-of-magnitude estimates; knowledge of their precise values is not necessary for understanding the conceptual points considered here. Later in this section we show how to relate these parameters back to nucleon-nucleon forces.

In the picture by Lane, Thomas, and Wigner,[21] for $W_p\cong\epsilon_n$, i.e., for energies of the compound states in the neighborhood of the single-particle energy ϵ_n, only terms $m=m'=n$ in the width γ_p of Eq. (17.1) are important, i.e., only the $(a_{on}{}^p)^2$ are large if W_p lies in the neighborhood of ϵ_n. We now make these criteria more quantitative. For simplicity, we consider the scattering of S-wave neutrons in the region of well-separated levels where the γ_p can be taken to be real.

We write the amplitude for compound-elastic scattering, defined by

$$S_\alpha{}^{ce}\equiv S_\alpha-\langle S_\alpha\rangle_{Av}, \tag{27}$$

in dispersion theory, using the fact that we choose $\tilde{V}$ so that $\tilde{S}_\alpha=\langle S_\alpha\rangle_{Av}$. Then,

$$S_\alpha{}^{ce}=S_\alpha-\tilde{S}_\alpha=\frac{e^{-2ikR}}{2k}\left\{\sum_p\frac{\gamma_p}{W_p-E}-\sum_m\frac{\Gamma_m}{\tilde{E}_m-E}\right\}, \tag{27.1}$$

where the second term in brackets gives the scattering amplitude in the complex well, and $\tilde{E}_m$ is the eigenvalue in the complex well, which may be taken to have the form, Eq. (9.13), $\tilde{E}_m=\epsilon_m-i\beta_m/2-iW$. (We are considering the complex well to be a square well here, although generalization to the case where it varies with r is easy.) From our definition, $\langle S_\alpha{}^{ce}\rangle_{Av}$, the average of

[20] F. L. Friedman and V. F. Weisskopf, *Niels Bohr and the Development of Physics* (Pergamon Press, London, 1955), p. 134.

[21] Lane, Thomas, and Wigner, Phys. Rev. **98**, 693 (1955).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

S^{ce}, is zero. Hence, the average of the quantity in brackets must vanish.[22] The intervals over which we average are always small compared with the widths of the single-particle resonances, so the second term in brackets need not be averaged. We can write

$$\left\{\sum_p \frac{\gamma_p}{W_p-E}\right\}_{\text{Av}} = P(E)+iQ(E) = \sum_m \frac{\Gamma_m}{\tilde{E}_m-E}, \quad (28)$$

where P and Q are real. We find

$$P(E)=\sum_m \frac{(\epsilon_m-E)\Gamma_m}{(\epsilon_m-E)^2+(W+\beta_m/2)^2}, \quad (28.1)$$

and

$$Q(E)=\sum_m \frac{\Gamma_m(W+\beta_m/2)}{(\epsilon_m-E)^2+(W+\beta_m/2)^2},$$

where Γ_m is assumed to be real, which implies specialization to the low-energy region. We cannot express $P(E)$ more simply in terms of γ_p, but the imaginary part of the sum $Q(E)$ can be put into a more useful form.

The average $\langle F(E)\rangle_{\text{Av}}$ of a quantity $F(E)$ is defined by

$$\langle F(E)\rangle_{\text{Av}} = \int_{-\infty}^{\infty} \rho(E-E')F(E')dE'. \quad (29)$$

The weighting function ρ employed by Feshbach, Porter, and Weisskopf[9] was a square one,

$$\rho(x)=\begin{cases}0, & x<-I/2\\ 1/I, & -I/2<x<I/2,\\ 0, & x>I/2\end{cases} \quad (29.1)$$

so that

$$\langle F(E)\rangle_{\text{Av}} = \frac{1}{I}\int_{E-I/2}^{E+I/2} F(E')dE'.$$

With such integrals, one has end effects coming from resonances where ϵ_p lies near either $E-I/2$ or $E+I/2$, and these must be disposed of, somewhat inelegantly. Physical results must be independent of the precise form of weighting function employed, as long as it is not an unreasonable one, and we find it convenient to use

$$\rho(E-E')=\frac{I}{\pi}\frac{1}{(E-E')^2+I^2},$$

in which case

$$\left\{\sum_p \frac{\gamma_p}{W_p-E}\right\}_{\text{Av}} = \frac{I}{\pi}\int_{-\infty}^{\infty}\sum_p \frac{\gamma_p}{W_p-E'}\frac{1}{(E-E')^2+I^2}dE'$$

$$=\sum_p \frac{\gamma_p}{W_p-E-iI}, \quad (30)$$

where the integral can be evaluated by contour integration. Quite generally, averaging with this weighting function a quantity $F(E)$ which has poles only in the lower half of the complex E plane (such functions are often called R functions following Wigner), one finds

$$\langle F(E)\rangle_{\text{Av}} = F(E+iI), \quad (31)$$

where I corresponds to the interval over which the average is carried out. We always assume this interva to contain many resonances p.

We find, then,

$$Q(E)=\text{Im}\sum_p \frac{\gamma_p}{W_p-E-iI}. \quad (32)$$

By defining the average density of the levels in the region of interest by $1/D$ (where $1/D=N$ gives the number of levels per unit energy, which is assumed to be large), one can convert the sum in Eq. (32) into an integral, obtaining

$$Q(E)=\frac{1}{D}\int_{-\infty}^{\infty}\frac{\bar{\gamma}I}{(\epsilon_p-E)^2+I^2}d\epsilon_p=\frac{\pi\bar{\gamma}}{D}, \quad (32.1)$$

where we have neglected α_p compared with I, and where $\bar{\gamma}$ is the average width. The function $\pi\bar{\gamma}/D$ is called the "strength function." Our assumption of well-separated levels is equivalent to $\pi\bar{\gamma}/D\ll 1$.

Finally, we find

$$\pi\frac{\bar{\gamma}}{D}=\text{Im}\sum_m \frac{\Gamma_m}{\tilde{E}_m-E}, \quad (32.2)$$

when we make the average value of the imaginary part of the quantity in brackets, Eq. (27.1), vanish, using $\langle S^{ce}\rangle_{\text{Av}}=0$. In the special case where the energy E is in the neighborhood of ϵ_n, the main contribution to the sum on the right-hand side comes from the nth term provided that the single-particle resonances are sufficiently far apart. Then,

$$\pi\frac{\bar{\gamma}}{D}\cong\frac{\Gamma_n(W+\beta_n/2)}{(\epsilon_n-E)^2+(W+\beta_n/2)^2}, \quad E\cong\epsilon_n. \quad (33)$$

The strength function, for constant W, Γ_n, and β_n, is thus of the Lorentz form. The contribution of the other levels is at least of order $W/\Delta\epsilon$ smaller,[23] where $\Delta\epsilon$ is

[22] In the low-energy region, where the k-dependent factor $k^{-1}\exp(-2ikR)$ and the penetrabilities in the widths vary rapidly with bombarding energy, these factors should be divided out before the average is taken. This is done explicitly in the next section.

[23] Since the sum in Eq. (32.2) relates only to the single-particle well, one can carry out the calculation of the sum explicity for special cases, such as a square well, and verify that if $\epsilon\cong\epsilon_n$, the sum of terms $m\neq n$, contributes in order $W/\Delta\epsilon$ (see Appendix A, reference 16).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

the distance between single-particle levels of the *same angular momentum*, i.e., of the same channel α. This distance is of the order of U, the real part of $\tilde{V}$.

The validity of the picture depends on $W \ll U$, which is also one of the criteria for the existence of marked resonances in $\tilde{V}$. In the range of separated compound-nucleus resonances, the condition $kR \ll 1$ is fulfilled, as appears later, and then the maxima in the strength function of resonance shape given by Eq. (33) are simply related to the total cross section σ_T, as can be seen by using the optical relation

$$\sigma_T = (4\pi/k)\ \mathrm{Im}\ S_\alpha. \tag{34}$$

Hence

$$\langle \sigma_T \rangle_{Av} = \frac{4\pi}{k}\ \mathrm{Im}\langle S_\alpha \rangle_{Av} = \frac{4\pi}{k}\ \mathrm{Im}\ \tilde{S}_\alpha \cong \frac{2\pi}{k^2}\pi\frac{\bar{\gamma}}{D}, \tag{34.1}$$

which we have obtained from Eq. (32.2).

This situation just discussed is illustrated in Fig. 6, where the distance between levels $\epsilon_{n'} - \epsilon_n$ is assumed to be much greater than W, so that the picture of Lane, Thomas, and Wigner is valid. The $4s$ and $5s$ single-particle levels are shown on the left-hand side, and it is illustrated on the right-hand side how—although each single-particle level is split up into many compound states—there is an appreciable probability of finding them only within an energy region of width $\sim W$ bracketing the single-particle energy.[24] Hence, although the interaction $V(\mathbf{r},\xi)$ is strong enough to change the wave function completely so as to split the single particle level $\tilde{\psi}_n$ into thousands or millions of compound states, it is not strong enough to mix the single-particle level $\tilde{\psi}_{n'}$, which is a distance of order $\epsilon_{n'} - \epsilon_n$ away, into the compound state in the neighborhood of ϵ_n with appreciable probability.

We might ask how well the condition $(\epsilon_{n'} - \epsilon_n) \gg W$ is fulfilled in the actual physical case, taking the parameters of Feshbach, Porter, and Weisskopf[9] for $\tilde{V}$.

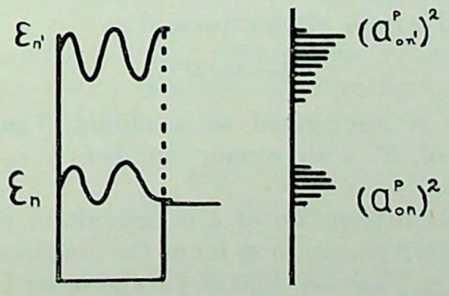


FIG. 6. On the left, S-wave single-particle resonances in the complex well are shown; on the right, the behavior of the square of the expansion coefficients. Behavior of the strength function is essentially the same as this.

By way of example, consider the nucleus A = 160 where the $4s$ single-particle resonance occurs near zero energy experimentally, as in Fig. 6. The occurrence of the $4s$ resonance requires $K_n R = (7/2)\pi$, where K_n is the wave number for a particle of zero energy, measured from the bottom of the well. The $5s$ resonance occurs for $K_{n'} R = (9/2)\pi$. Hence, $\epsilon_{n'}/\epsilon_n = 81/49$, where ϵ_n and $\epsilon_{n'}$ are now measured from the bottom of the well. For a well depth of 42 Mev, $\epsilon_{n'} - \epsilon_n \equiv \Delta\epsilon = 27$ Mev. Thus there is no doubt that $W \ll \Delta\epsilon$ is the low-energy region, where W is of order 1–2 Mev.[25]

Although the single-particle width Γ_n is split up among thousands or millions of compound states, some features of the single-particle resonance remain. In particular, since to a very good approximation,

$$\gamma_p \cong (a_{on}{}^p)^2 \Gamma_n, \quad E \cong \epsilon_n,$$

and since it follows from the completeness of our two representations connected by the coefficients $a_{jm}{}^p$ that $\sum_p (a_{on}{}^p)^2 = 1$, then

$$\sum_p{}' \gamma_p \cong \Gamma_n,$$

where the prime on the sum indicates that it is extended only over compound states in the neighborhood of ϵ_n. In other words, the single-particle width is split up among many levels, but the sum of widths of these many levels is just equal to the single-particle width.

At this stage we have an understanding of why the compound resonances p are so narrow without resorting to the classical picture which explains this in terms of the energy being shared among all of the nucleons. The interaction $V(\mathbf{r},\xi)$ is sufficiently strong so as to mix many states $\chi_j(\xi)\tilde{\psi}_m{}^j(\mathbf{r})$ into the state $\Phi^{(p)}$. Since

$$\sum_{j,m} (a_{jm}{}^p)^2 = 1,$$

following from the normalization of the $\Phi^{(p)}$, and since there are many terms which contribute in this sum, then $(a_{on}{}^p)^2 \ll 1$ for any given p and the width of the compound resonance is much less than that of the single-particle one.

Nonetheless, the interaction $V(\mathbf{r},\xi)$ is not strong enough to break down the underlying single-particle structure completely and, in particular, it does not mix different single-particle levels n and n' into the same $\Phi^{(p)}$, but only spreads the single-particle resonance locally. Consequently, it is not surprising that some of the single-particle features remain in experiments carried out with wide beams.

For completeness, we remark that the width of the strength function is given essentially by W as shown

[24] We have made arguments only for the case of separated levels where the $a_{on}{}^p$ are essentially real, and therefore our arguments do not apply in detail to the higher single-particle resonance n'. However, it is clear from them that the $(a_{on'}{}^p)^2$ has a spread of order W, where W is the imaginary part of $\tilde{V}$ (which is velocity-dependent) at the excitation energy $\epsilon_{n'}$.

[25] The discussion of the contribution of terms $m \neq n$ in Eq. (32.2) was carried out taking the same W for all terms m, contrary to the spirit of the other development where we think of $\tilde{V}$ as a velocity-dependent potential, or, equivalently, a nonlocal potential $\tilde{V}(\mathbf{r}',\mathbf{r})$, in which case the W's entering into the widths of different single-particle levels are quite different. However, the sum of terms $m \neq n$ must be the same in either method.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

in Eq. (33) (although this equation is a reasonable approximation only when E is in the neighborhood of ϵ_n), and not by the square root of the second moment of the perturbing potential, as Lane, Thomas, and Wigner[21] indicated might be the case. Bloch[26] has pointed out, using some rough quantitative estimates, that one should expect the root-mean-square moment to be much larger than the width owing to large contributions from the wings of the strength function which are weighted heavily. Use of the form Eq. (33) for all energies would lead to an infinite second moment which would contradict the finite expression for it in terms of the potential,[27] so that this form is clearly not valid far from the resonance energy.

3. Low-Energy Behavior of the Scattering[28]

At low energies, where $(\pi\bar{\gamma}/D)\ll 1$ and $kR\ll 1$, the interpretation of average quantities is especially simple. Equation (34.1) shows that the average cross section is proportional to the strength function in this region. The maxima in the strength function indicated by Eq. (33) would be difficult to observe experimentally by measuring the strength function at various energies for a given nucleus because the single-particle resonance is broad, of width $2W+\beta_n$, and even if the same experimental technique could be used over this broad energy interval, the interpretation would become difficult, both because waves of higher angular momenta would be mixed into the experiments and because our simple consideration—which applied only to the region of well-separated resonances where $kR\ll 1$—would no longer be valid. Consequently, the strength function is usually measured at low energies for different nuclei and plotted as a function of A. In this way a variation in the quantity $E-\epsilon_n$ is obtained which depends on KR, where K is the wave number measured from the bottom of the well. Some experimental data, plotted in this way, are shown in Fig. 7.[29] The theoretical curve is not computed here from dispersion formalism but is obtained from

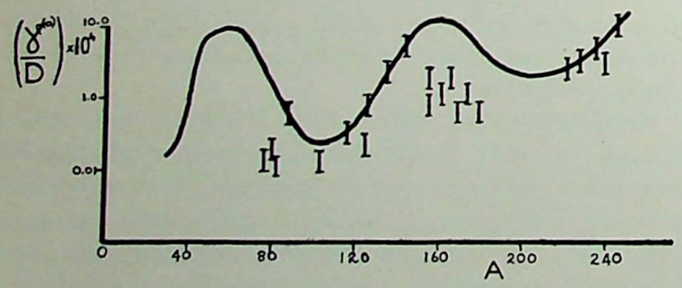


FIG. 7. Experimental data for the strength function. The solid line indicates typical theoretical predictions for a complex-well model. A measured width γ_p is reduced to a $\gamma^{(0)}$ by using $\gamma^{(0)}=\gamma_p(k_0/k)$, where k_0 is taken to be the wave number for a neutron of energy 1 ev. A more refined theory[29] takes into account nuclear deformation, and the resulting agreement between theory and experiment is then better.

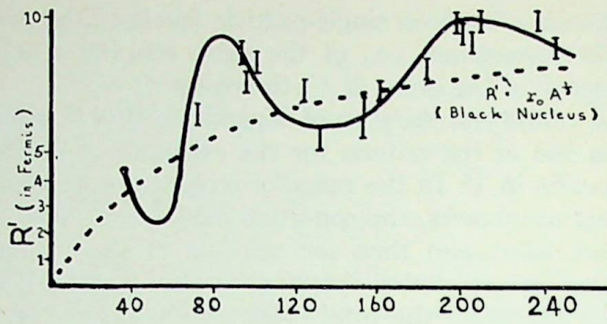


FIG. 8. Experimental data compared with theoretical predictions for R' (solid line) from the complex-well model. A more refined theory[31] takes into account nuclear deformations, and the resulting agreement between theory and experiment is then better.

$$\pi\frac{\bar{\gamma}}{D}=2k\ \mathrm{Im}\ \tilde{S}_\alpha, \tag{34.2}$$

[see Eq. (34.1)], where $\tilde{S}_\alpha$ is calculated directly by integrating the Schrödinger equation with potential $\tilde{V}$. Nevertheless, our formulas are useful in understanding the qualitative behavior in the resonance region.

Historically the first striking observations of the nonmonotonic behavior of average cross sections with changing A referred to measurement[30] of a quantity sometimes called the "potential scattering." To obtain this quantity, we expand in k, considering k to be small, but must carry the expansion one step further, i.e., we express the average cross section $\langle\sigma_T\rangle_{\mathrm{Av}}$ as

$$\langle\sigma_T\rangle_{\mathrm{Av}}=(a/k)+b, \tag{35}$$

which can be done at low energies where terms of higher order in k are negligible. The first term on the right-hand side is the familiar $1/v$ term in the cross section which is related to the strength function and which was already discussed in Eqs. (34.1) and (34.2), and the second term is often expressed as

$$b=4\pi(R')^2, \tag{35.1}$$

where R' is interpreted as a radius. The observed behavior of R' with atomic number A is shown in Fig. 8.[31]

A simple description of b is difficult to give in the Kapur-Peierls theory in so far as the imaginary parts of the width γ_p [see Eq. (23) ff.] are of order k compared with the real parts, and these must therefore be taken into account. It is consequently more straightforward to give this description in the Wigner-Eisenbud for-

[26] C. Bloch, Nuclear Phys. 3, 137 (1957).
[27] The second moment is clearly finite for well-behaved potentials which are used here.
[28] Many of the results in this section were first derived in the R-matrix formalism by R. G. Thomas, Phys. Rev. 97, 224 (1955).
[29] Hughes, Zimmerman, and Chrien, Phys. Rev. Letters 1, 461 (1958).
[30] Fields, Russell, Sachs, and Wattenberg, Phys. Rev. 71, 308 (1947).
[31] Seth, Hughes, Zimmerman, and Garth, Phys. Rev. 110, 692 (1958).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

malism. We sketch this briefly, following the treatment of Lane and Lynn,[32] and refer to the review article of Lane and Thomas[1] for a detailed discussion.

The expansion is actually in powers of $\rho=kR$, as in Eq. (9.5) ff., because ρ is a good expansion parameter at the energies of interest. To make this explicit, we write

$$2ik\langle S\rangle_{Av}=e^{-2i\rho}-1+2i\rho e^{-2i\rho}\langle R_{KP}\rangle_{Av}, \tag{36}$$

where

$$R_{KP}=\sum_p \chi_p/(W_p-E), \tag{36.1}$$

R_{KP} being the Kapur-Peierls R function. (It has poles only in the lower half of the complex plane.) χ_p is the reduced width,

$$\chi_p=(2\rho)^{-1}\gamma_p. \tag{36.2}$$

The relation between R_{KP} and the Wigner R function R_W is easily obtained by looking at the scattering matrix in the two theories. It is

$$R_{KP}=R_W/(1-i\rho R_W). \tag{37}$$

In averaging R_{KP} by adding iI to E, we find that

$$\langle R_{KP}\rangle_{Av}=\bar{R}_W/(1-i\rho\bar{R}_W), \tag{37.1}$$

(see pp. 306–309 of Lane and Thomas), where

$$\bar{R}_W=R_W{}^{\infty}+i\pi s_W \tag{37.2}$$

with $R_W{}^{\infty}$ and s_W now real, and given by

$$s_W=\frac{\bar{\chi}^{(0)}}{D}, \quad R_W{}^{\infty}=\int\frac{s_W(E')}{E'-E}dE'. \tag{37.3}$$

Here $\bar{\chi}^{(0)}$ is the average of $\chi_p{}^{(0)}$, the width with the Wigner-Eisenbud boundary conditions [see Eq. (23)]. The expansion to any required order is now easily made.

$$\frac{4\pi}{k}\,\mathrm{Im}\langle S_\alpha\rangle_{Av}=\frac{4\pi\rho}{k^2}(\pi s_W)+\frac{4\pi\rho^2}{k^2}\left[(1-R_W{}^{\infty})^2-(\pi s_W)^2\right]+O(\rho^3). \tag{38}$$

The left-hand side of Eq. (38) is equal to $\langle\sigma_T\rangle_{Av}$ by the relation Eq. (34.1). We now can identify easily the a and b of Eq. (35); thus,

$$a=\frac{4\pi\rho}{k^2}(\pi s_W),$$

and

$$b=4\pi\rho^2[(1-R_W{}^{\infty})^2-(\pi s_W)^2]. \tag{38.1}$$

To lowest order in ρ, $2\rho\pi s_W=\pi\bar{\gamma}/D$ and, retaining only a, we see that Eq. (38) reduces to our earlier expression, Eq. (34.1).

This derivation is independent of any particular model, although the model introduced in the last section is helpful in understanding the qualitative behavior of s_W and $R_W{}^{\infty}$ in the actual physical case. We have already indicated the resonance-type behavior in the strength function in Eq. (33). We see immediately from Eq. (37.3) that to the extent that the strength function is symmetrical about its maximum, $R_W{}^{\infty}$ is zero at this maximum which occurs at the real part of the single-particle resonance energy. Further, $R_W{}^{\infty}$ becomes negative on one side of the maximum and positive on the other so that one can understand the qualitative behavior of R' shown in Fig. 8.

Theoretical curves, such as those shown by the solid lines in Figs. 7 and 8, are obtained by calculating $\tilde{S}_\alpha$ and then using $\langle S_\alpha\rangle_{Av}=\tilde{S}_\alpha$.

In past work the term $(\pi s_W)^2$ has often been neglected. However, near the single-particle resonance at A = 50, s_W is appreciable, as R. K. Seth has pointed out,[33] and it is vital to include the term $(\pi s_W)^2$ in analyzing data near A = 50. Otherwise $(1-R_W{}^{\infty})$ comes out to be too large.

This treatment illustrates the type of problem in which the Wigner-Eisenbud formalism is useful. It is advantageous here to deal with the real quantities of this theory in which the energy dependence is explicit.

Reverting to the Kapur-Peierls formalism, we can show the relation between the average cross section for forming a compound nucleus and the W in this energy region, to lowest order in k. The former is defined as

$$\langle\sigma_c\rangle_{Av}\equiv\langle\sigma_T\rangle_{Av}-\sigma_{se},$$

where σ_{se} is the cross section for shape-elastic scattering,

$$\sigma_{se}=4\pi|\langle S_\alpha\rangle_{Av}|^2=4\pi|\tilde{S}_\alpha|^2.$$

Use of the relation between $\langle\sigma_T\rangle_{Av}$ and Im $\tilde{S}_\alpha$ given in Eq. (34.1) results in

$$\langle\sigma_c\rangle_{Av}=4\pi\left\{\frac{1}{k}\,\mathrm{Im}\,\tilde{S}_\alpha-|\tilde{S}_\alpha|^2\right\}. \tag{38.2}$$

By using the resonance expansion for $\tilde{S}_\alpha$ [obtained from Eq. (7) by replacing E_m by $\bar{E}_m$] and remembering that Γ_n is of order kR, we find, to lowest order in kR,

$$\langle\sigma_c\rangle_{Av}\cong\frac{4\pi}{k^2}\sum_m\frac{W\Gamma_m/2}{(E-\epsilon_m)^2+W^2}, \tag{38.3}$$

where we see that the shape-elastic scattering does not contribute in this order. To interpret this formula we consider $\psi_\alpha(\mathbf{r})$, the solution of the Schrödinger equation in the complex well. Since we are dealing with S waves we can expand ϕ_α, the radial part of ψ_α multiplied by r, in terms of eigenstates $\tilde{\phi}_m$. The coefficient has already been derived in Eq. (4.3). Thus

$$\phi_\alpha=\frac{h^2}{2M}e^{-ikR}\sum_m\frac{\tilde{\phi}_m(R)\tilde{\phi}_m(r)}{\bar{E}_m-E}. \tag{38.4}$$

[32] A. M. Lane and J. E. Lynn, "The widths and spacings of resonance levels," Harwell Report T/R 2210 (1957, unpublished).

[33] R. K. Seth, Optical Model Conference, Tallahassee, Florida, March, 1959.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

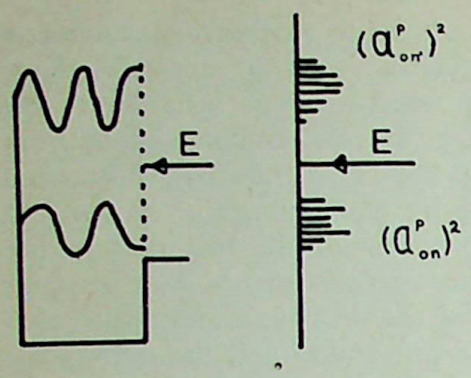


FIG. 9. Illustration of the position of the neutron energy with respect to the single-particle resonances in our "simple case of scattering."

This formula clearly shows the resonance behavior of the probability

$$\mathcal{P}_\alpha \equiv \int_0^R |\phi_\alpha(r)|^2 dr = \left(\frac{\hbar^2}{2M}\right)^2 \sum_m \frac{[\tilde{\phi}_m(R)]^2}{(E-\epsilon_m)^2+W^2}, \tag{39}$$

of finding a particle inside the nucleus whose wave function has unit amplitude asymptotically. Here we have dropped β_m which is of order kR smaller than W. It is clear that

$$\langle \sigma_c \rangle_{\text{Av}} = \frac{8\pi M}{\hbar^2 k} \mathcal{P}_\alpha W. \tag{40}$$

In other words, the average cross section for compound-nucleus formation just defined is proportional to the absorption W times the probability $\mathcal{P}_\alpha$ of finding the nucleon inside the nucleus. In a somewhat classical way, one might interpret W as the absorption per unit nucleon in the nucleus.

4. A Simple Case of Scattering

Much of the relationship of our detailed picture, where we make a complete description in terms of the $\Phi^{(p)}$ to the simplified complex-well model, can be understood by considering an especially simple case[34,35] in which the energy of the incident particle falls between single-particle resonance energies. Let us suppose, referring to Fig. 6, that the energy E of the incident particle lies midway between ϵ_n and $\epsilon_{n'}$, as in Fig. 9. The amplitude for elastic scattering coming from the lower group of levels is

$$(S_\alpha)_n = \frac{e^{-2ikR}}{2k} \sum_p \frac{(a_{on}{}^p)^2 \Gamma_n}{W_p - E} - \frac{\sin kR}{k} e^{-ikR}, \tag{41}$$

where the lower suffix n on S_α indicates that this is the S_α that would result if only the lower group of levels existed. We have here used the model developed in the last section and dropped the $a_{on'}{}^p$ in the widths γ_p referring to the set of compound-nucleus levels lying around ϵ_n.

The W_p are spread a distance W about ϵ_n. Consequently, if $E-\epsilon_n \gg W$, then a good first approximation is to set $E-W_p$ equal to $E-\epsilon_n$ in the denominator of Eq. (41). Evaluation of the sum over p gives in this approximation,

$$(S_\alpha)_n = \frac{e^{-2ikR}}{2k} \frac{\Gamma_n}{\epsilon_n - E} - \frac{\sin kR}{k} e^{-ikR}, \tag{42}$$

which is just the contribution to S_α from the nth resonance in the complex-well picture if we neglect W compared with $E-\epsilon_n$ in the denominator.

Therefore, if the spread of the γ_p about the single-particle resonance energy is small, then the whole set of them act together just as the single-particle resonance. This is related to the principle of spectroscopic stability, from which it follows that—although the wave functions may be drastically changed by the perturbation—certain simple features come out just as in the unperturbed problem.

The main oversimplification is in the neglect of the widths γ_p which lie close to E, i.e., the γ_p are not so well concentrated around ϵ_n and $\epsilon_{n'}$ as we have shown. However, our approximations are good in so far as the "shape-elastic" scattering is concerned. Our neglect of the near levels has simply resulted in our dropping the compound-elastic scattering. We now make the development more precise in such a way that the near levels can also be treated. Such a treatment is clearly necessary in order to consider general E which may be in the region of a single-particle resonance.

5. Expression of the Scattering in the Green's Function Formalism

In defining $\tilde{V}(r)$, it is convenient to express S^{ce} in terms of the Green's function as was done by Bloch.[36] Actually, Bloch's elegant formalism is much more general than that we use, and by specialization to specific representations he can obtain either the Kapur-Peierls or Wigner-Eisenbud expression. However, our simpler formalism is adequate for illustrating the points considered here. We again specialize our consideration to the scattering of S-wave neutrons; generalization of our results to other cases is obvious.

We use the notation

$$\Phi^{(p)}(\mathbf{r},\xi) = |p\rangle = \langle p|, \quad \chi_j(\xi)\tilde{\psi}_m{}^j(\mathbf{r}) = |jm\rangle = \langle jm|, \tag{43}$$

and

$$\langle p|jm\rangle = \int_0^R \Phi^{(p)}(\mathbf{r},\xi)\chi_j(\xi)\tilde{\psi}_m{}^j(\mathbf{r}) d^3r d^3\xi = a_{jm}{}^p.$$

No distinction is made between $|p\rangle$ and $\langle p|$ since expansion and orthogonality conditions in this theory involve only the wave function and not its complex conjugate, aside from the trivial changes in functions of angles mentioned in part 2 of this section.

[34] J. Bowcock, Proc. Phys. Soc. (London) **A70**, 515 (1957).
[35] G. E. Brown and C. T. De Dominicis, Proc. Phys. Soc. (London) **A70**, 668 (1957).
[36] C. Bloch, Nuclear Phys. 4, 503 (1957).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

We then have

$$S_\alpha{}^{ce}=\frac{e^{-2ikR}}{2k}\left\{\sum_p\frac{\gamma_p}{W_p-E}-\sum_m\frac{\Gamma_m}{\tilde{E}_m-E}\right\},$$

$$=\frac{h^2}{2M}e^{-2ikR}\sum_{m,m'}\left\{\sum_p\langle p|\,om\rangle\langle p|\,om'\rangle\left\langle p\left|\frac{1}{H-E}\right|p\right\rangle\right.$$

$$\left.-\left\langle om\left|\frac{1}{\tilde{H}-E}\right|om'\right\rangle\right\}\tilde{\phi}_m(R)\tilde{\phi}_{m'}(R), \quad (44)$$

$$=\frac{h^2}{2M}e^{-2ikR}\sum_{m,m'}\left\{\left\langle om\left|\frac{1}{H-E}\right|om'\right\rangle\right.$$

$$\left.-\left\langle om\left|\frac{1}{\tilde{H}-E}\right|om'\right\rangle\right\}\tilde{\phi}_m(R)\tilde{\phi}_{m'}(R),$$

where $\tilde{H}$ is obtained from H by replacing $V(\mathbf{r},\xi)$ by $\tilde{V}$, i.e.,

$$H=\tilde{H}+(V(\mathbf{r},\xi)-\tilde{V}(r))=\tilde{H}+\delta V. \quad (44.1)$$

By using the identity

$$\frac{1}{H-E}=\frac{1}{\tilde{H}-E}-\frac{1}{\tilde{H}-E}(V-\tilde{V})\frac{1}{\tilde{H}-E}$$

$$+\frac{1}{\tilde{H}-E}(V-\tilde{V})\frac{1}{H-E}(V-\tilde{V})\frac{1}{\tilde{H}-E}, \quad (45)$$

which is easily checked by multiplying both sides of the equation by $H-E$, we obtain

$$S_\alpha{}^{ce}=-\frac{h^2}{2M}e^{-2ikR}\sum_{m,m'}\frac{1}{\tilde{E}_m-E}\frac{1}{\tilde{E}_{m'}-E}$$

$$\times\left\{\langle om|\,V-\tilde{V}|\,om'\rangle-\left\langle om\left|(V-\tilde{V})\frac{1}{H-E}\right.\right.\right.$$

$$\left.\left.\left.\times(V-\tilde{V})\right|om'\right\rangle\right\}\tilde{\phi}_m(R)\tilde{\phi}_{m'}(R). \quad (46)$$

At this stage, by using Eq. (38.4) we obtain $S_\alpha{}^{ce}$ in a convenient form,

$$S_\alpha{}^{ce}=-\frac{2M}{h^2}\left\{\langle o\alpha|\,V-\tilde{V}|\,o\alpha\rangle\right.$$

$$\left.-\left\langle o\alpha\left|(V-\tilde{V})\frac{1}{H-E}(V-\tilde{V})\right|o\alpha\right\rangle\right\}, \quad (47)$$

where ψ_α is the scattering state defined by Eq. (11) and the discussion following it. Further expansion of $1/(H-E)$ by iteration with Eq. (45) would give the Born expansion for $S_\alpha{}^{ce}$, i.e., the expansion in successive powers of the perturbing potential δV. Equation (47) can be derived directly from scattering theory without recourse to dispersion theory.

Insertion of the unit operator $\sum_p|\,p\rangle\langle p\,|$ in the second term on the right-hand sides of Eqs. (46) and (47) allows replacement of $(H-E)^{-1}$ by $(W_p-E)^{-1}$, indicating that all of the poles of this term lie in the lower half of the complex energy plane. Thus we can average this term with respect to energy by replacing E by $E+iI$, following the procedure of Eqs. (30) and (31).

$$\langle S_\alpha{}^{ce}(E)\rangle_{Av}=S_\alpha{}^{ce}(E+iI). \quad (48)$$

The complex potential $\tilde{V}(r)$ is defined by the condition $\tilde{S}_\alpha(E)=\langle S_\alpha(E)\rangle_{Av}$, or by $\langle S_\alpha{}^{ce}(E)\rangle_{Av}=0$. This can be accomplished by requiring

$$\langle o|\,V-\tilde{V}|\,o\rangle$$

$$-\left\langle o\left|(V-\tilde{V})\frac{1}{H-E-iI}(V-\tilde{V})\right|o\right\rangle=0. \quad (49)$$

One can therefore define an optical potential $\tilde{V}(r)$ which reproduces the average scattering phase. At this stage, not very much has been proved because the formalism does not make it clear that $\tilde{V}$ does not vary rapidly and possibly nonmonotonically with energy; its usefulness empirically comes from the fact that it varies slowly and regularly with energy and with atomic number A. We now go as far as we can towards demonstrating that $\tilde{V}$ has these characteristics. In order to do this, we make the connection between the parameters of $\tilde{V}$ and the nucleon-nucleon potentials.

6. Perturbation Theory of the First Kind

In the perturbation theory to be developed, we will have to carry out an expansion in $\delta V=V(\mathbf{r},\xi)-\tilde{V}(r)$ since we wish to obtain expressions for quantities which do not contain the complicated $\Phi^{(p)}$. The smallness of the widths of the $\Phi^{(p)}$ compared with those of the single-particle states $\tilde{\psi}_m$ which, together with χ_j we employ as unperturbed wave functions, indicates that a large number of terms are necessary in the expansion Eq. (14) of $\Phi^{(p)}$ in terms of $\chi_j(\xi)\tilde{\psi}_m{}^j(\mathbf{r})$. Thus, an expansion of $\Phi^{(p)}$ in perturbation theory, starting from $\chi_0(\xi)\tilde{\psi}_m{}^j(\mathbf{r})$ as zero-order function, would converge very slowly, if at all. However, the perturbation expansions of the average phase involve less stringent criteria for convergence than those for the wave functions as might be suspected from the physical arguments of the first part of this section. We now formulate these criteria, following the development of Brown, De Dominicis, and Langer.[37]

$\mathfrak{W}$ is defined as

$$\tilde{V}=\bar{V}-\mathfrak{W}, \quad (49.1)$$

where

$$\bar{V}\equiv\langle o|\,V|\,o\rangle. \quad (49.2)$$

[37] Brown, De Dominicis, and Langer, Ann. Phys. (N. Y.) **6**, 209 (1959).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

Our defining equation, Eq. (49), for $\tilde{V}$ becomes

$$\left\langle o\left|(V-\bar{V}+\mathfrak{W})\left[1-\frac{1}{H-E-iI}(V-\bar{V}+\mathfrak{W})\right]\right|o\right\rangle=0. \tag{49.3}$$

Equation (49.3) can be explicitly solved for $\mathfrak{W}$.[38] For brevity,

$$e=H-E-iI, \tag{49.4}$$

and

$$\bar{e}=\bar{H}-E-iI=e-(V-\bar{V}).$$

Then Eq. (49.3) can be written

$$\begin{aligned}&\left\langle o\left|(e-\bar{e}+\mathfrak{W})\left[1-\frac{1}{e}(e-\bar{e}+\mathfrak{W})\right]\right|o\right\rangle=0,\\ &=\left\langle o\left|(e-\bar{e}+\mathfrak{W})\frac{1}{e}(\bar{e}-\mathfrak{W})\right|o\right\rangle,\\ &=\left\langle o\left|(e-\bar{e})\frac{1}{e}\bar{e}\right|o\right\rangle-\left\langle o\left|(e-\bar{e})\frac{1}{e}\right|o\right\rangle\mathfrak{W}\\ &\qquad+\mathfrak{W}\left\langle o\left|\frac{1}{e}(\bar{e}-\mathfrak{W}\right|o\right\rangle.\end{aligned} \tag{49.5}$$

Since $H_\xi|O\rangle=0$, the factor $(\bar{e}-\mathfrak{W})$ in the final term commutes with $|o\rangle$ and can be taken out to the right. Equations (45)–(49) imply

$$\left\langle o\left|\frac{1}{e}\right|o\right\rangle=\left\langle o\left|\frac{1}{\bar{e}-\mathfrak{W}}\right|o\right\rangle, \tag{49.6}$$

and from this we see that the final term in Eq. (49.5) is equal to $\mathfrak{W}$. We can then solve the equation for $\mathfrak{W}$, obtaining

$$\mathfrak{W}=\frac{1}{1-\left\langle o\left|(V-\bar{V})\frac{1}{e}\right|o\right\rangle}\left\langle o\left|(V-\bar{V})\frac{1}{e}(V-\bar{V})\right|o\right\rangle, \tag{49.7}$$

where we have added to the final factor the term

$$\left\langle o\left|(e-\bar{e})\frac{1}{e}e\right|o\right\rangle,$$

for symmetry. This term is zero, since $\langle o|e-\bar{e}|o\rangle=0$. Equation (49.7) is equivalent to

$$\mathfrak{W}=\left\langle o\left|(V-\bar{V})\frac{1}{e-\Lambda_o(V-\bar{V})}(V-\bar{V})\right|o\right\rangle, \tag{50}$$

where

$$\Lambda_o=|o\rangle\langle o|, \tag{50.1}$$

as can be checked by expanding in powers of $(V-\bar{V})1/e$. One can derive this expression without recourse to perturbation theory as an intermediate step.[39,40] For purposes of perturbation theory it is more useful to think of the denominator of Eq. (50) as $\bar{e}+(1-\Lambda_o)(V-\bar{V})$, since the expansion is in powers of $(1-\Lambda_o)(V-\bar{V})$.

In working with Eq. (50) it is advantageous to use a representation for the extra particle which is diagonal in $T+\bar{V}$ rather than in $T+\tilde{V}$. The importance of considering $\mathfrak{W}$ comes from the fact that it contains the entire imaginary part of $\tilde{V}$; its contribution to the real part of $\tilde{V}$ would be important only in case one was trying to calculate this quantity fairly accurately. For the simple case in which $\tilde{V}$ is a square well, changing the imaginary part of it by iW does not change th eeigenfunctions but only shifts the resonance energies by iW. Our new energy eigenvalues are, therefore,

$$E_m=\epsilon_m-i\beta_m/2. \tag{50.2}$$

We now investigate the eigenfunctions and eigenvalues of the operator

$$\mathfrak{H}=H-\Lambda_o(V-\bar{V}). \tag{51}$$

We label them by Ω_p and W_p', i.e.,

$$\mathfrak{H}\Omega_p=W_p'\Omega_p. \tag{51.1}$$

Now, $-\Lambda_o(V-\bar{V})$ can be considered as a small perturbation since the Ω_p are not very different from the $\Phi^{(p)}$. Thus, we choose

$$\Omega_p=\Phi^{(p)}+\delta\Omega_p, \tag{52}$$

and

$$\delta\Omega_p=\textstyle\sum_n a_p{}^n\Phi^{(n)}.$$

To lowest order

$$W_p'=W_p. \tag{52.1}$$

The use of Green's theorem as in Eq. (16) and the assumption[41] that Ω_p satisfies the same boundary condition as $\Phi^{(p)}$, gives the result

$$\begin{aligned}&\int_0^R(\Phi^{(n)}\mathfrak{H}\Omega_p-\Omega_p\mathfrak{H}\Phi^{(n)})d^3\xi d^3r=0\\ &=(W_p-W_n)a_p{}^n+\sum_m\langle p|(V-\bar{V})|om\rangle\langle om|n\rangle,\end{aligned} \tag{53}$$

where $W_p'=W_p$ and $\Omega_p=\Phi^{(p)}$ on the right-hand side since we wish to compute $a_p{}^n$ only to lowest order. By employing

$$\begin{aligned}\langle p|V-\bar{V}|om\rangle&=\langle p|H-\bar{H}|om\rangle\\ &=(W_p-\epsilon_m+i\beta_m/2)\langle p|om\rangle,\end{aligned} \tag{54}$$

[38] C. T. De Dominicis, J. phys. radium **19**, 1 (1958).

[39] J. S. Langer, thesis, University of Birmingham, 1958.

[40] H. Feshbach, Ann. Rev. Nuclear Sci. 8, 44 (1958).

[41] Actually, the functions Ω_p obey slightly different boundary conditions from the $\Phi^{(n)}$, because, as is clear from Eq. (51), the Ω_p contain no components χ_0, these being projected out by the operator Λ_0. However, such components of $\Phi^{(n)}$ do not contribute to the left-hand side of Eq. (53) because they are orthogonal to Ω_p, and the equality given there holds.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

we obtain

$$a_p{}^n=\sum_m \frac{(W_p-\epsilon_m+i\beta_m/2)}{W_n-W_p}\langle p|\,om\rangle\langle om|\,n\rangle. \quad (55)$$

For order-of-magnitude estimates,

$$\langle p|\,om\rangle\langle om|\,n\rangle\beta_m\sim\langle p|\,om\rangle^2\beta_m\sim\gamma_p,$$

and on the giant resonance

$$W_p-\epsilon_m\lesssim W.$$

As a result the $a_p{}^n$ are large only if $(W_p-W_n)\sim W\gamma_p/\Gamma_m$ and, therefore, because of the smallness of γ_p, only a few of the neighboring Φ's are mixed into each Ω by the perturbation.

We can expand the Ω's in terms of the χ's and the ψ's just as with the $\Phi^{(p)}$'s. Then

$$\Omega_p=\sum_{j,m} b_{jm}{}^p\chi_j(\xi)\bar{\psi}_m{}^j(\mathbf{r}). \quad (55.1)$$

One important difference between the right-hand side of Eq. (55.1) and that of the expansion of the $\Phi^{(p)}$, Eq. (14), is that the right-hand side of Eq. (55.1) contains no terms with $j=0$. These have been projected out by the operator Λ_o, so that the eigenfunctions of $\mathfrak{IC}$ form a complete set in the space orthogonal to χ_0.

Since the eigenvalues and eigenfunctions of $\mathfrak{IC}$ differ only slightly from those of H, aside from the fact that the former do not contain components in $|o\rangle$, our arguments about the distribution of the $a_{jm}{}^p$ apply equally well to the distribution of the $b_{jm}{}^p$. However, it is important to retain the Λ_o in the denominator of Eq. (50), i.e., not to approximate the Ω_p by the $\Phi^{(p)}$,[42] because this would allow states $|o\rangle$ to occur as intermediate states in the expansion which would give large spurious contributions. Retaining the Λ_o is related to the need to eliminate unlinked clusters in expansions of the Brueckner type.

We are now ready to discuss the criteria for perturbation theory. We can express the matrix element as

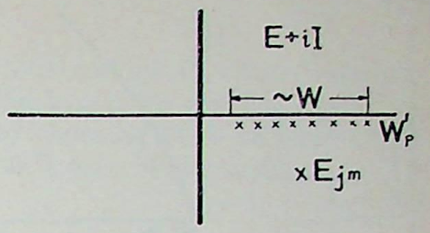
$$\langle on|\,\mathfrak{W}(E)\,|on\rangle=\langle on|\,V-\bar{V}\,|p'\rangle \times\frac{1}{W_{p'}-E-iI}\langle p'|\,V-\bar{V}\,|on\rangle, \quad (56)$$

where we label the eigenfunctions of $\mathfrak{IC}$ by $|p'\rangle$. In a perturbation expansion of the type

$$\mathfrak{W}(E)=\left\langle o\left|(V-\bar{V})\frac{1}{\bar{H}-E-iI}\sum_{n=0}^{\infty}(-1)^n \times\left[(1-\Lambda_0)(V-\bar{V})\frac{1}{\bar{H}-E-iI}\right]^n(V-\bar{V})\right|o\right\rangle, \quad (57)$$

[42] This was pointed out to the author by Dr. A. M. Lane. The partial summation of Bloch[26] takes this into account properly.

FIG. 10. Positions of poles in the complex-energy plane. The scattering amplitude, averaged over energy interval I, is obtained by taking its value for the complex energy $E+iI$.



the lowest-order term of the expansion can be expressed as

$$\langle on|\,\mathfrak{W}(E)\,|on\rangle\cong\sum_{j,m}\frac{\langle on|\,V-\bar{V}\,|jm\rangle\langle jm|\,V-\bar{V}\,|on\rangle}{E_{jm}-E-iI}. \quad (58)$$

The same result can be obtained by replacing $|p'\rangle$ by $\sum_{j,m}|\,jm\rangle\langle jm|\,p'\rangle$ in Eq. (56) and then approximating W_p' by E_{jm}. This approximation is justified if I is large, because the $\langle jm|\,p'\rangle$ are large only if $W_p'-\epsilon_{jm}\lesssim W$ by the arguments of the preceding section. The situation is illustrated in Fig. 10. In the complex energy plane, the state $|\,jm\rangle$ has an appreciable probability of being found in states $|p'\rangle$ distributed over a region of width W. In going from Eq. (56) to the approximation Eq. (58), we are approximating the distances $E+iI-W_p'$ in the energy plane by the common distance $E+iI-E_{jm}$. This is justified if $|E+iI-E_{jm}|\gg W$, W giving the spread of the W_p' as indicated in Fig. 10. In the worst case for the fulfillment of this criterion, in which $E=\epsilon_{jm}$, which is illustrated in Fig. 10, we can satisfy this criterion by making I sufficiently large, so we must have $I\gg W$.

This could be phrased in the following way: Calculating only the average phase means that we can evaluate it for the imaginary energy $\mathcal{E}=E+iI$. The approximation of taking only the first term in the expansion is clearly better, the larger $\mathcal{E}-E_{jm}$. In the simple case of scattering considered earlier (where we actually approximated at a different stage), a large value for the distance between $\mathcal{E}$ and the single-particle resonance was obtained by taking E, the real part of $\mathcal{E}$, to be far away. The same result can be achieved by pushing $\mathcal{E}$ up in the complex plane, which means averaging over larger intervals.

We have thus formulated mathematically the physical considerations of part 1 of this section. There it was seen that the time spent in the nucleus by the particles corresponding to the average phase decreased as the interval I, over which the average was carried out, increased. This development shows that the larger I is, the faster the series Eq. (57) converges. This series in powers of δV can be interpreted as an expansion in the number of collisions the incident particle makes with the target nucleons. The larger I is, the more quickly the particle passes over the nucleus and the fewer the collisions it makes.

The expansion of $\mathfrak{W}$ in this way has been carried out by Bloch.[36] After expansion, he makes assumptions

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

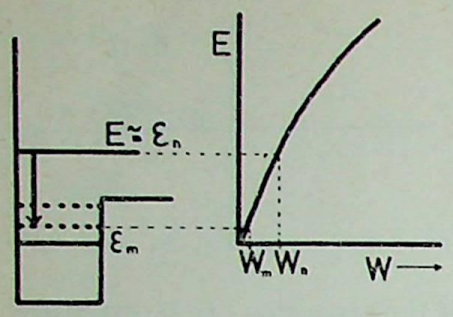


FIG. 11. Variation of absorption W with energy of the single-particle excitation.

about the randomness of signs of the $\langle kl|\delta V|jm\rangle$. We do not, however, make such assumptions, and show later that they are not justified. The expansion is, however, convergent without these assumptions provided I is sufficiently large.

It might appear that the condition $I \gg W$ is difficult to achieve. The decisive point here is that the W to be employed is the W appropriate to the state $|jm\rangle$ which corresponds to the nucleus excited by energy ϵ_j and a single particle of energy roughly $E-\epsilon_j$ in the well. This is illustrated in Fig. 11. From our earlier developments, we can assume energy conservation to within the width of the excitation $|jm\rangle$. Thus $E \cong \epsilon_j + \epsilon_m$. The width of the state $|j\rangle$, which is a true compound state, can be neglected relative to that of the single-particle excitation $|m\rangle$. But the W corresponding to state $|m\rangle$ is much smaller than that corresponding to state $|n\rangle$ because the former is far down in the well, and W is a rapidly increasing function of energy, as shown in Fig. 11. Taking the quadratic dependence of W given by simple theories, as discussed later,

$$W_m = \left(\frac{E-\epsilon_j}{E}\right)^2 W_n, \tag{59}$$

where we have put a lower suffix m on the W on the left-hand side to indicate that it refers to the state $|m\rangle$, and we now denote the W referring to state $|n\rangle$ by W_n. In other words, the width of the state into which the single-particle excitation decays is less than the width of the original single-particle excitation.

Usually, the state j is a highly excited one since the number of states per Mev available increases exponentially with excitation,[43] and in this case $W_m \ll W_n$. Of course, the more highly excited states also tend to have a more complicated structure so that the matrix elements for excitation of the A particles become smaller. In the next part of this section it is shown that the ratio $(E-\epsilon_j)/E$ tends to be $\sim 1/3$ or, $W_m/W_n \sim 1/9$. However, in the case of easily deformable nuclei, there are large matrix elements to low-lying states j, the collective ones.

From the foregoing, we see that the inequalities

$$W_m < I < W_n, \tag{60}$$

can usually be satisfied, although W_m is smaller than W_n only by a numerical factor and not by orders of magnitude. When these inequalities are satisfied, the average can be carried out over an interval which is large enough for the convergence of perturbation theory but small enough so that one can still obtain information about the shape of the single-particle resonance of width $\sim W_n$. In fact, neglect of the variation in the $|n\rangle$ and other quantities in the various averages constrains us to average over a distance $< W_n$.

The same arguments apply, with slight modification, to the case of inelastic processes going into a low-excited final state. Here the transition amplitude is

$$S_{\alpha\alpha',j} = -\frac{2M}{\hbar^2}\Big\{\langle j\alpha'|V-\tilde{V}|o\alpha\rangle - \Big\langle j\alpha'\Big|(V-\tilde{V})\frac{1}{H-E}(V-\tilde{V})\Big|o\alpha\Big\rangle\Big\},$$

$$= -\frac{\hbar^2}{2M}e^{-i(k+k_j)R}\sum_{m,m'}\frac{1}{\tilde{E}_m - E}\frac{1}{\tilde{E}_{m'}-E}$$

$$\times\Big\{\langle jm'|V-\tilde{V}|om\rangle - \Big\langle jm'\Big|(V-\tilde{V})\frac{1}{H-E}$$

$$\times(V-\tilde{V})\Big|om\Big\rangle\Big\}\tilde{\phi}_m(R)\tilde{\phi}_{m'}{}^{j}(R),$$

where j is assumed to be a low-excited state. Perturbation theory can be used to calculate the $\langle S_{\alpha\alpha',j}\rangle_{\text{Av}}$, since the states that m and m' decay into are, in general, of substantially lower energy and consequently have a smaller width. This gives the result, in the first approximation,

$$\langle S_{\alpha\alpha',j}\rangle_{\text{Av}} \cong -\frac{2M}{\hbar^2}\langle j\alpha'|V-\bar{V}|o\alpha\rangle, \tag{62}$$

which has been used extensively in calculating direct interaction processes. This is just the same result as follows from lowest-order Born approximation, i.e., keeping only the lowest term in δV with initial and final states distorted by the complex well. Such a weak interaction picture is only applicable to the average phase and then only if it is averaged over a large energy interval.

7. Perturbation Theory of the Second Kind

Relation (62) is an important one in that it justifies the use of perturbation theory—in the usual terminology, Born approximation with distorted waves—in direct interaction calculations. We found also, in part 6, that the average elastic scattering amplitude was reproduced by $\tilde{V} = \bar{V} - \mathfrak{W}$ with $\mathfrak{W}$ given to lowest order by Eq. (58). This latter relation is interesting but not well adapted for the calculation of $\mathfrak{W}$ since it still

[43] The energy of excitation is limited since the initially excited particle does not drop into a state already occupied. This is inconsistent with our neglect of antisymmetrization, but this does not seem to be an essential difficulty.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

involves the highly complicated nuclear states $\chi_j(\xi)$. We therefore develop a perturbation expansion of the type used by Bruckner, Eden, and Francis[44] which essentially relates quantities back to shell-model states. This expansion gives an expression for $\mathcal{W}$ which is practical for calculation; however, to satisfy the criterion for its validity, one must average over a large energy interval as we shall see.

In this treatment the unperturbed Hamiltonian is

$$\bar{H}=\sum_{i=0}^{A}(T_i+V_i)-\tfrac{1}{2}\sum_{i,j=0}^{A}\bar{V}_{ij}, \tag{63}$$

with eigenstates $|\bar{q}\rangle$ such that

$$\bar{H}|\bar{q}\rangle=E_{\bar{q}}|\bar{q}\rangle. \tag{63.1}$$

Here V_i is the self-consistent potential defined in the state $|\bar{q}\rangle$ felt by the ith particle, as in the Hartree-Fock theory, and

$$\bar{V}_{ij}=\langle\bar{q}|V_{ij}|\bar{q}\rangle. \tag{63.2}$$

The summations cover the incident particle ($i=0$) as well as the A particles in the nucleus. Thus, the A+1 particles are here treated symmetrically. The c number, $\frac{1}{2}\sum_{i,j}\bar{V}_{ij}$, has been subtracted to ensure that $\langle\bar{q}|H|\bar{q}\rangle$ is equal to the energy in the Hartree-Fock approximation. The crucial question in using the $|\bar{q}\rangle$ as zero-order functions relates to the extent of the spread of the strength function $\langle\bar{q}|p\rangle^2$ in energy. This can be obtained from

$$\frac{\pi}{D}\{\langle\bar{q}|p\rangle^2\}_{Av}=\mathrm{Im}\left\{\left\langle\bar{q}\left|\frac{1}{H-E}\right|\bar{q}\right\rangle\right\}_{Av}$$
$$=\mathrm{Im}\left\langle\bar{q}\left|\frac{1}{\tilde{H}-E}\right|\bar{q}\right\rangle, \tag{64}$$

where $\tilde{H}$ is a comparison Hamiltonian which has been introduced as an artifice to calculate this strength function.[45] Its significance is seen later. It is given by

$$\tilde{H}=\bar{H}-\mathcal{W}, \tag{64.1}$$

where $\mathcal{W}$ is then defined by

$$0=\left\langle\bar{q}\left|\mathcal{W}\right|\bar{q}\right\rangle$$
$$-\left\langle\bar{q}\left|(V-\bar{V}+\mathcal{W})\frac{1}{H-E-iI}(V-\bar{V}+\mathcal{W})\right|\bar{q}\right\rangle, \tag{64.2}$$

which guarantees satisfaction of Eq. (64). [See the similar development, Eqs. (44)–(49).] This equation is formally similar to Eq. (56), and by the same procedure as employed to solve the latter, we can find that Eq. (64.2) can be satisfied by choosing $\mathcal{W}$ to be diagonal in the $|\bar{q}\rangle$ representation, with matrix elements

$$\langle\bar{q}|\mathcal{W}|\bar{q}\rangle$$
$$=\left\langle\bar{q}\left|(H-\bar{H})\frac{1}{H-\Lambda_{\bar{q}}(V-\bar{V})-E-iI}(H-\bar{H})\right|\bar{q}\right\rangle, \tag{65}$$

where

$$\Lambda_{\bar{q}}=|\bar{q}\rangle\langle\bar{q}|. \tag{65.1}$$

In evaluating $\langle\bar{q}|\mathcal{W}|\bar{q}\rangle$ by lowest-order perturbation theory which now means replacing

$$\{H-\Lambda_{\bar{q}}(H-\bar{H})-E-iI\}^{-1}$$

by

$$\sum_{\bar{s}\neq\bar{q}}|\bar{s}\rangle\{E_{\bar{s}}-E-iI\}^{-1}\langle\bar{s}|,$$

we find that

$$\mathcal{W}=\sum_{i<j=0}^{A}\mathcal{W}_{ij}, \tag{65.3}$$

with

$$\langle\bar{q}|\mathcal{W}_{ij}|\bar{q}\rangle=\sum_{\bar{s}\neq\bar{q}}\langle\bar{q}|V_{ij}|\bar{s}\rangle\frac{1}{E_{\bar{s}}-E-iI}\langle\bar{s}|V_{ij}|\bar{q}\rangle. \tag{65.4}$$

The half-width of the strength function $\pi\{\langle\bar{q}|p\rangle^2\}_{Av}/D$ is given by

$$\sum_{i<j=0}^{A}\mathrm{Im}\langle\bar{q}|\mathcal{W}_{ij}|\bar{q}\rangle, \tag{65.5}$$

if we neglect the natural width of the state $|\bar{q}\rangle$.

If we identify $|\bar{q}\rangle$ with the state in which A+1 particles form a state, wherein the A levels from the bottom of the well are filled except for one hole in the state we label by k' and in which two particles are excited in states labeled by k and l, then it is easy to identify the various parts of the width:

1. $\mathrm{Im}\sum_{j=0}^{A}\langle\bar{q}|\mathcal{W}_{kj}|\bar{q}\rangle$ and $\mathrm{Im}\sum_{i=0}^{A}\langle\bar{q}|\mathcal{W}_{li}|\bar{q}\rangle$,

correspond to the widths for the particles k and l to interact with the other particles so that the two-particle excitation, in perturbation theory language, decays into a three-particle excitation. (One should omit $j=k'$ and $i=k'$ in the sums.)

2. $\mathrm{Im}\sum_{j}{}'\langle\bar{q}|\mathcal{W}_{k'j}|\bar{q}\rangle$ corresponds to the width for the decay of the shell model state with a hole in it, i.e., the width for absorption of the hole. (The prime indicates that the term $j=k'$ should be omitted. But $j=k$ and l should, for completeness, be included in the sum.)

By using the approximate symmetry of holes and particles[46] near the Fermi surface E_F, we find that the

[44] Brueckner, Eden, and Francis, Phys. Rev. **100**, 891 (1955). See also M. Cini and S. Fubini, Nuovo cimento **2**, 75 (1955) and A. M. Lane and C. F. Wandel, Phys. Rev. **98**, 1524 (1955).

[45] We use the symbols $\tilde{H}$ and $\mathcal{W}$ because, although they stand now for different quantities than previously, there is a close analogy between these and the previous ones.

[46] Calculations by E. P. Pendlebury, analogous to those of Lane and Wandel,[44] indicate that this symmetry holds only near the Fermi surface; the absorption for holes then rises more steeply than that for particles (private communication).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

total width of $|\bar{q}\rangle$ is

$$W_{\bar{q}}=W_k+W_l+W_{k'}$$

$$\cong\left\{\frac{(\epsilon_k-E_F)^2+(\epsilon_l-E_F)^2+(E_F-\epsilon_{k'})^2}{(\epsilon_n-E_F)^2}\right\}W_n, \quad (66)$$

from the simple theories referred to previously. We label the width of the original single-particle excitation $|n\rangle$ by W_n as previously, and we employ the W's to denote the imaginary parts of the $\mathcal{W}_{ij}$.

The width of our original single-particle excitation is given in the lowest approximation by

$$W_n=\mathrm{Im}\sum_{\bar{q}}$$

$$\frac{\langle on|\sum_{i=1}^{A}V_{oi}-\sum\bar{V}_{oi}|\bar{q}\rangle\langle\bar{q}|\sum V_{oi}-\sum\bar{V}_{oi}|on\rangle}{E_{\bar{q}}-E-iI}. \quad (67)$$

The largest phase space is available when the energy is equally distributed among the particles and the hole so that W_k, W_l, and $W_{k'}$ are each $\sim(1/9)W_n$. Since $W_k+W_l+W_{k'}<W$, we can again satisfy

$$W_{\bar{q}}<I<W_n, \quad (68)$$

and employ perturbation theory, although all three of these quantities are of the same order of magnitude, so that the average must be carried out over an interval I that is a good fraction of the width of the single-particle resonance.

In case $|o\rangle$ can be described as a shell-model state, i.e., as an antisymmetrized wave function of independent particles moving in a well, or can be obtained from such a state by successive application of the two-body interaction, the A+1 particles can be treated symmetrically, and the identity of the particles can be taken into account easily. By using such an iteration in which the evaluation is carried out starting from the Fermi sea, which is assumed to be a reasonable approximation for the interior of a large nucleus, Brueckner[47] finds that the value of W is increased considerably over that given by the simple theory and introduces both a term linear in energy and a constant term in the dependence of W. This would mean that the criterion, Eq. (67), might not be satisfied by any I since the sum $W_k+W_l+W_{k'}$ might be as large as W_n. Whereas Brueckner's calculation brings out the sensitivity of W to correlations in the ground-state wave function near the Fermi surface, there are good indications that these cannot be calculated in perturbation theory, i.e., that the ground-state wave function $\chi_o(\xi)$ cannot be obtained in the region of the Fermi surface by iterating the shell-model state.[48] In fact, the parts of these correlations responsible for the pairing forces tend to make the Fermi surface stiffer against perturbations, working in the opposite direction from the effects calculated by Brueckner.[47] Consequently, a consistent evaluation of the corrections to the lowest-order theory undoubtedly gives results smaller than Brueckner's calculation.

In the discussion following Eq. (59) we promised to return to the question of easily deformable nuclei where there are large matrix elements from χ_o to low-lying rotational states so that the initial single-particle excitation has a high probability of decaying into an excitation where the initial particle has only slightly less energy. This state, therefore, has a width almost as large as the initial width. In this case, one should separate out the strong transitions to the low-lying collective states which, together with the initial channel, we term the "chosen channels" and treat them separately in a system of coupled equations as has been done by Sano, Yoshida, and Terasawa,[49] (references to extensive earlier work by Yoshida and others are given in this article) and by Chase, Wilets, and Edmonds.[50]

In fact, the preceding development can be generalized so that $V-\bar{V}$ is a matrix between chosen channels, $|o\rangle$ becoming the space of the chosen channels and Λ_o excluding these from occurring in intermediate states, essentially reproducing Yoshida's formalism. This generalization is within the spirit of the optical model where the distortion provided by the central potential is supposed to represent the average effects from the great number of channels and correspondingly large number of degrees of freedom which cannot be conveniently treated in detail.

8. Special Models

It would not be possible to evaluate $\mathcal{W}$ from either Eq. (56) or Eq. (58) since the matrix elements $\langle on|V-\bar{V}|p\rangle$ or $\langle on|V-\bar{V}|jm\rangle$ cannot be calculated owing to the complexity of the $\Phi^{(p)}(\mathbf{r},\xi)$ and the $\chi_j(\xi)$. It would be difficult to evaluate $\mathcal{W}$ even from the equations of the last section since these would involve using wave functions in a complex well of finite extent. Consequently, somewhat idealized models have been introduced to carry out this evaluation and we now discuss their relationship to the development here.

The interior of a large nucleus has been represented as a Fermi gas. In this case, the absorption, Eq. (67), is approximated by

$$W_n=\frac{\pi}{D}\{\langle on|\sum V_{oi}-\sum\bar{V}_{oi}|\bar{q}\rangle^2\}_{\mathrm{Av}}, \quad (69)$$

where the Fermi sea $|o\rangle$ is taken as an antisymmetrized product function of plane waves, $|n\rangle$ is a plane wave of energy above the top of the Fermi sea, and $|\bar{q}\rangle$ is a

[47] K. A. Brueckner, Phys. Rev. **103**, 172 (1956).
[48] Bohr, Mottelson, and Pines, Phys. Rev. **110**, 936 (1958).
[49] Sano, Yoshida, and Terasawa, Nuclear Phys. **6**, 20 (1958).
[50] Chase, Wilets, and Edmonds, Phys. Rev. **110**, 1080 (1958).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

state of the same energy as $|on\rangle$ in which a second particle has been lifted above the sea.

In such a model it is immaterial how large the interval I over which the average is taken in calculating $\mathfrak{W}$ or W. No features are left in the model capable of giving rapid variations in the quantities calculated, so that the scattering amplitude is a smoothly varying function of energy and the average phase, for reasonable intervals, is very close to the actual phase. Consequently, one does not need to carry the iI in the energy denominator in the actual evaluations. However, in establishing the validity of Eq. (67), it was necessary to carry the iI in the compound-state picture.

A second model, which is not so useful in obtaining numerical values, but is instructive, has been used by Wigner.[51] He assumes that levels $|jm\rangle$ are evenly spaced in energy and that the matrix elements $\langle jm|\delta V|kl\rangle$, where δV is the perturbation, are all equal in absolute value but of random sign as long as $E_{jm}-E_{kl}$ is less than a certain quantity (in the Wigner formalism the energy eigenvalues are real) and zero for greater energy differences. He then finds, under certain conditions which are probably satisfied in the nuclear case, a Lorentzian form, i.e., the shape Eq. (33), for the strength function in the region of the single-particle resonance. This implies, according to Eq. (33), that W is constant with respect to energy.

A third model, used by Bloch,[36] is similar to Wigner's except that the point at which he makes his assumption of random signs and equality of matrix elements is in the perturbation expansion, Eq. (57). This means that only the even powers of the expansion contribute. His analysis shows that the largest number of nonvanishing terms comes from the lowest-order term, corresponding to Eq. (58). However, such assumptions of random signs are unjustified in the actual physical case, and for the case of a large nucleus, when approximated by a Fermi sea, it is found that the third term in the series,[52] which would be zero under the above assumptions, actually contributes about as much to the imaginary part of the potential as the second term. Furthermore, the second term appears to be much less predominant when the exclusion principle is taken into account since the second-order processes are severely inhibited.

Once one has adopted a simplified model, any mechanism that might be present in the actual nucleus to give rapid variations in the parameters has been dropped, and consequently, the result is insensitive to the interval over which one averages since the actual phase is very nearly the same as the average phase. However, the inclusion of iI when the compound-state energies W_p were present was necessary to establish the relevance of these models to the calculation of the average phase.

It seems reasonable that at high bombarding energies, where the widths of the compound states are much greater than their spacings, the actual scattering amplitude varies smoothly with energy and that fluctuations are small. Consequently, one might expect the actual amplitude to be the same as the average one here, and again, the calculated amplitude would be insensitive to the distances over which one averaged. This would explain why our formulas for the average amplitude have the same form as those of Watson and collaborators for higher energy scattering even though no average over energy is carried out in their work.

However, in the intermediate region where fluctuations may still be important and in the low-energy region, it is necessary to carry out the averages over energy as we have done.

We have not discussed the fluctuation scattering in much detail. This is because it is not possible to evaluate this directly; since the corresponding particles stay in the nucleus a long time, any expansion in terms of successive collisions is bound to fail. In fact, because of their long duration in the nucleus, these particles would seem to correspond to those described by the original extreme compound-nucleus model. Consequently, it is reasonable to neglect phase relations here on the average, and this results in angular distributions symmetrical about 90°. There is always some ambiguity about adding in the fluctuation scattering, but if one can calculate the average scattering amplitude so as to obtain the first term on the right-hand side of Eq. (26), one obtains a lower limit on the cross section for the relevant process. This is especially useful in inelastic scattering where similar considerations apply. An application of this is indicated in the next section. Furthermore, if the above assumptions are justified, the fluctuation scattering is rather structureless and simply tends to make the minima in the shape-elastic scattering less pronounced without altering the general picture.

IV. DIRECT INTERACTION IN THE DIPOLE PHOTOEFFECT AND IN RADIATIVE CAPTURE

We treat one case in which the incident and emerging particles are different, namely, that in which one of the particles is a photon. As discussed in Sec. I, a giant resonance is observed in the absorption of photons by complex nuclei in the region of ~15 Mev for medium weight nuclei. Collective models give a natural explanation of the absorption mechanism but not of the number of fast protons emerging which is orders of magnitude in excess of the predictions of the statistical theory. Direct interactions are necessary to account for them. This does not mean that emission of fast protons is the dominant decay process; this constitutes only ~1% of the total decay processes in heavy nuclei where such decays are strongly inhibited by the Coulomb barrier. However, this is several orders of magnitude above the predictions of 10^{-4} or 10^{-5} for the probability given by the statistical theory. The

[51] E. P. Wigner, Ann. Math. **62**, 548 (1955).
[52] L. Verlet and J. Gavoret, Nuovo cimento **10**, 505 (1958).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

inverse (p,γ) processes violate the predictions of the statistical theory in just as striking a fashion.[53,54]

Whereas this theory would predict a strong decrease in the cross section with increasing energy or increasing atomic number A, the experimental results of Cohen[55] show no such decreases for proton energies in excess of 5 Mev. This is an added indication that direct processes play an important role.

Wilkinson[14] has proposed a model for the (γ,p) reaction which incorporates the physical features that follow from the detailed theory. He views the absorption of the γ-ray by one of the nucleons as leading first to an excited state of the single particle of width W in the complex well. This is just the single-particle excitation discussed earlier. One then associates with this level the natural width Γ for escape from the nucleus and the width $2W$ for absorption into compound states. Consequently, the proportion $\Gamma/(\Gamma+2W)$ escape, carrying off essentially the full energy of the γ-ray. This reproduces the observed order-of-magnitude of fast protons when averaged over the relevant single-particle levels.

We now extend the formalism to treat the absorption of radiation following Brown and Levinger.[56] This extension is similar to that carried out in the framework of the continuum theory of nuclear reactions by Peaslee.[57] The wave function of the system in which γ-rays are absorbed consists only of outgoing waves and is described by

$$\Psi(\mathbf{r},\xi)=\sum_{\alpha',j} S_{\gamma\alpha',j}\chi_j(\xi)\psi_{\alpha'}{}^{j+}(\mathbf{r}), \quad r>R, \tag{70}$$

where $S_{\gamma\alpha',j}$ is the amplitude for γ-ray absorption with emission of the nucleon into channel α', leaving the residual nucleus in state j. For $r<R$ we have

$$\Psi=\Psi_o(\mathbf{r},\xi)+\textstyle\sum_p a_p\Phi^{(p)}(\mathbf{r},\xi), \tag{71}$$

where the $\Phi^{(p)}$ are the compound-state functions, Eq. (13). Here Ψ_o is the wave function of the initial nucleus (it being understood that a γ-ray is present). It can often be adequately represented as a shell-model wave function. The coordinate of the particle to be emitted is labeled by $\mathbf{r}$ whereas those of all the other A$-$1 particles are labeled by ξ. Thus, although there is a symmetry between the A particles, we choose to describe the absorption by a particular particle for simplicity, and our description consequently singles out this particle. The interaction with the radiation field (electric field $\mathcal{E}$ along the z axis normalized to a flux of one photon per cm^2 sec) is described by H_I where

$$H_I=e\mathcal{E}_z=e(2\pi\hbar\omega)^{\frac{1}{2}}z \tag{72}$$

on restricting the consideration to the dipole interaction. Since the coupling to the radiation field is weak, the a_p's are needed only to first order which can easily be done by requiring Ψ to satisfy the Schrödinger equation to that order. We have

$$H_I\Psi_o=(E-H)\textstyle\sum a_p\Phi^{(p)}=\sum a_p(E-W_p)\Phi^{(p)}. \tag{73}$$

By multiplying on the left by $\Phi^{(p)}$ and integrating, we find

$$a_p=\langle p|H_I|o\rangle/(E-W_p), \tag{73.1}$$

as expected from perturbation theory. Here we use $|o\rangle$ for Ψ_o; E is the full energy of the system. We now measure it from the ground state of the initial nucleus, $E=\hbar\omega$, where $\hbar\omega$ is the γ-ray energy.[58]

We easily obtain $S_{\gamma\alpha',j}$ by equating Ψ from Eqs. (71) and (73) at $r=R$, multiplying by $\chi_j(\xi)\Theta_\alpha(\theta,\varphi)$ and integrating over $d^3\xi$ and $d\Omega$.

$$S_{\gamma\alpha',j}=\frac{1}{\phi_{\alpha'}{}^{j+}(R)}\sum_p\int_0^R d^3\xi\int d\Omega\Phi^{(p)}(\xi,\mathbf{R})\chi_j(\xi)\times\Theta_\alpha(\theta,\varphi)\frac{\langle p|H_I|o\rangle}{W_p-\hbar\omega}. \tag{74}$$

From the expansion, Eq. (14), we find

$$S_{\gamma\alpha',j}=\frac{1}{\phi_{\alpha'}{}^{j+}(R)}\sum_{p,m}\tilde{\phi}_m{}^j(R)\langle jm|p\rangle\frac{\langle p|H_I|o\rangle}{W_p-\hbar\omega}, \tag{74.1}$$

This gives a cross section

$$\sigma_{\gamma\alpha',j}=\frac{\hbar k_j}{Mc}|S_{\gamma\alpha',j}|^2. \tag{75}$$

Once again we are concerned with the average scattering amplitude $\langle S_{\gamma\alpha',j}\rangle_{\text{Av}}$ which can be obtained from Eq. (74.1) by replacing $\hbar\omega$ by $\hbar\omega+iI$ in the denominator, i.e.,

$$\langle S_{\gamma\alpha',j}\rangle_{\text{Av}}=\frac{1}{\phi_{\alpha'}{}^{j+}(R)}\sum_{p,m}\tilde{\phi}_m{}^j(R)\langle jm|p\rangle\frac{\langle p|H_I|o\rangle}{W_p-E-iI}, \tag{76}$$

which in the equivalent Green's function description is

$$\langle S_{\gamma\alpha',j}\rangle_{\text{Av}}=\frac{1}{\phi_{\alpha'}{}^{j+}(R)}\sum_m\tilde{\phi}_m{}^j(R)\left(jm\left|\frac{1}{H-E-iI}H_I\right|o\right). \tag{76.1}$$

The equivalence of Eqs. (76) and (76.1) is easily shown by inserting the unit operator $|p\rangle\langle p|$ to the left of H_I in the latter. By employing an expansion similar to that, Eq. (45) for $(H-\hbar\omega-iI)^{-1}$, one has

[53] F. Beck, Nuclear Phys. 9, 140 (1958/9).
[54] A. M. Lane and J. E. Lynn, Nuclear Phys. 11, 646 (1959).
[55] B. L. Cohen, Phys. Rev. 100, 206 (1955).
[56] G. E. Brown and J. S. Levinger, Proc. Phys. Soc. (London) A71, 733 (1958).
[57] D. C. Peaslee, Phys. Rev. 88, 812 (1952).

[58] This means that the energy of the excited particle, in terms of single-particle excitations, is measured from the energy of the ground state of the original nucleus, whereas earlier, the E associated with the single particle in our discussion of elastic scattering was its energy at $r=\infty$.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

$$\langle S_{\gamma\alpha',j}\rangle_{\text{Av}}=\frac{1}{\phi_{\alpha'}{}^{i+}(R)}\sum_m \tilde{\phi}_m{}^{i}(R)\Big\{\frac{\langle jm|H_I|o\rangle}{\bar{E}_{jm}-\hbar\omega-iI}$$
$$-\frac{1}{\bar{E}_{jm}-\hbar\omega-iI}\frac{1}{\bar{E}_{kl}-\hbar\omega-iI}$$
$$\times\Big[\langle jm|V-\tilde{V}|kl\rangle$$
$$-\frac{\langle jm|V-\tilde{V}|p\rangle\langle p|V-\tilde{V}|kl\rangle}{W_p-\hbar\omega-iI}\Big]$$
$$\times\langle kl|H_I|o\rangle\Big\}, \quad (77)$$

where the label k refers to the excited states of the A$-$1 particles as does j. We now limit consideration to the case in which $\epsilon_{jm}\cong E$, i.e., the single particle takes off approximately the full energy of the γ-ray. The other case, in which $\epsilon_{jm}-E\gg W_m$ where W_m is the absorption relating to the single-particle state $|m\rangle$, can be treated just as our simple case of scattering, Sec. III, 4.

Perturbation theory here amounts to expanding the $(W_p-\hbar\omega-iI)^{-1}$ and the first term is obtained by replacing $|p\rangle(W_p-\hbar\omega-iI)^{-1}\langle p|$ by

$$\sum_{i,n}|in\rangle(E_{in}-\hbar\omega-iI)^{-1}\langle in|.$$

As discussed in Sec. III, 6, this is a good approximation because the states $|n\rangle$ are mainly at a lower excitation than the original $|l\rangle$ which are the states reached by the single particle absorbing approximately the full γ-ray energy, and consequently the spread of the $\langle p|in\rangle^2$ is small compared with W_l, the absorption pertaining to the state $|l\rangle$.

We can expect the first term on the right-hand side of Eq. (77) to be a good approximation to $\langle S_{\gamma\alpha',j}\rangle_{\text{Av}}$. We consider first the case $|jm\rangle=|kl\rangle$. In this case, it is useful to consider both terms in square brackets together. If it is remembered that the overlap between $|j\rangle$ and the part of Ψ_o relating to the ξ-particles is large only if the states of the ξ-particles are similar in both (the final factor $\langle kl|H_I|o\rangle$ ensures this will be as shown), and if $\tilde{V}$ is chosen so as to satisfy Eq. (49), then the term in square brackets here very nearly vanishes, any remaining effects coming from the small differences between $|j\rangle$ and the ξ-variable part of Ψ_o. Since the two terms in the square brackets are separately of order W_m, and since the denominators in front never become smaller than W_m+I, then if the two terms cancel to a good approximation, their contribution can be neglected. Estimates of the first term in square brackets for the case $|jm\rangle\neq|kl\rangle$ indicate that it is of order $(\bar{V}/\text{A})/(W_l+I)$ relative to the first term. Corrections for this term may have to be made in some cases. We do not pursue this further but assume that the first term gives a sufficiently good approximation. In any case, we have indicated that the series expansion converges for large enough I so that we could calculate further terms if necessary for good numerical results.

The first term on the right-hand side of Eq. (77) looks very much like the matrix element that would describe the photonuclear process for a particle in a complex well, since the states $|p\rangle$ have disappeared. Remnants of the fact that the process is really not a single-particle one remains however, in that states $|j\rangle$ are employed rather than the shell-model excited states which are just the states in which particles fill all but one of the A levels from the bottom of the well. To show how we can dispose of the details of the states $|j\rangle$, let us assume that the initial state Ψ_o is well represented by the shell-model state so that we can write

$$\Psi_o(\mathbf{r},\xi)=\Omega(\xi)\psi_{m'}(\mathbf{r}), \quad (78)$$

where $\psi_{m'}(\mathbf{r})$ is the bound state occupied initially by the r-particle. Then we can split the matrix element coming into the first term of Eq. (77) into

$$\langle jm|H_I|o\rangle=\langle m|H_I|m'\rangle\int_0^R\Omega(\xi)\chi_j(\xi)d^3\xi$$
$$=\langle m|H_I|m'\rangle Q_{\Omega j}, \quad (78.1)$$

defining $Q_{\Omega j}$. Let us assume that only one single-particle level contributes essentially to a given channel which is often a good approximation. The cross section arising from the first term on the right-hand side of Eq. (77) is

$$\sum_i\frac{\hbar k_j}{Mc}|\langle S_{\gamma\alpha',j}\rangle_{\text{Av}}|^2\cong\sum_i\Gamma_{\alpha'j}Q_{\Omega j}{}^2\frac{1}{\hbar c}\left|\frac{\langle m|H_I|m'\rangle}{\bar{E}_{jm}-\hbar\omega-iI}\right|^2, \quad (79)$$

where

$$\Gamma_{\alpha'j}\equiv\frac{\hbar k_j}{M}\left|\frac{1}{\phi_{\alpha'}{}^{i+}(R)}\right|^2|\tilde{\phi}_m{}^{i}(R)|^2, \quad (79.1)$$

is the full single-particle escape width taken at energy $\hbar\omega-\epsilon_j$ where ϵ_j is the excitation energy of the state j. From completeness, the probability of finding the state Ω somewhere in the excited states $|j\rangle$ is unity, i.e.,

$$\sum_j Q_{\Omega j}{}^2=1. \quad (79.2)$$

From the development in Sec. III, 7, the width of the distribution of ϵ_j about $|\epsilon_{m'}|$ is $W_{m'}$, where $W_{m'}$ can be interpreted as the width of the hole in state m'. Consequently, we see that the width of the total distribution of emitted fast particles is the sum of the width for the excited particle and of the width of the hole in the shell-model states.

These arguments do not add any new physical features to Wilkinson's model but show how this simple description relates to the many-body description.

Similar equations can be used to estimate the cross section for direct radiative capture of neutrons or

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

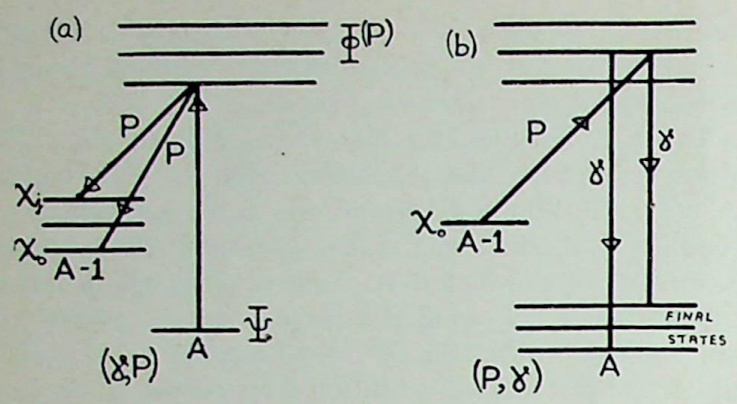


FIG. 12. Relation of the (γ,p) to (p,γ) reactions.

protons. An extensive description of this is contained in work by Lane and Lynn.[54]

The cross sections for (γ,p) and (p,γ) reactions can of course easily be related. We have the (γ,p) cross section by summing over all members of the configuration of the final $A-1$ nucleons (see Fig. 12). The (γ,p) reaction goes from the ground state to all final states j. The (p,γ) reaction may go from the ground state $\chi_o(\xi)$ to the various final states [see Fig. 12(b)] which contain an appreciable component of $\chi_o(\xi)$. (The wave function for the ξ-particles must remain the same since H_I is a function only of $\mathbf{r}$.) Further, the final state must contain an appreciable component of $\psi_{m'}(\mathbf{r})$, the relevant single-particle state in the well. From arguments similar to those already used the probability of finding $\chi_o\psi_{m'}$ somewhere in the ground or low-excited states is essentially unity. Thus, the cross section for the (p,γ) reaction, summed over the ground and low-excited states, is given by Eq. (79) with two minor changes: (i) we multiply by the usual factor k_γ^2/k_p^2 from detailed balance or dispersion theory where k_γ and k_p are γ-ray and proton wave numbers; (ii) we use $\Gamma_{\alpha' j}$ without the averaging implied in the sum over j in Eq. (79) since the incident particle now has a unique energy.

V. DISCUSSION OF OTHER FORMALISMS

Recently, a unified theory of nuclear reactions has been formulated by Feshbach,[59] and many of the relations between compound-nucleus parameters and those of the optical model have been derived by him, although he does not develop the perturbation theory described in Sec. III. Conceptually, his formalism has the advantage that it avoids the introduction of a joining radius. However, his formalism does not seem to be as convenient for making many of the arguments of Sec. III and Sec. IV as the one used here. In any case, there is a close connection between his formalism and the one employed here even though they appear quite different at first sight. We briefly demonstrate some of the relationship, trying to preserve most of Feshbach's notation so that one can easily compare our formulas with his papers.

[59] H. Feshbach, Ann. Phys. 5, 357 (1958).

Feshbach employs functions $\Phi_n(\mathbf{r},\xi)$ which are eigenfunctions of the operator $\mathcal{H}$, Eq. (51), and which are, consequently, just the Ω_p of Eq. (51). The Λ_o in Eq. (51) effectively projects off the incident channel so that the Φ_n form a complete set of states in the space orthogonal to that of the incident channel. Hence, one can write the solution of the Schrödinger equation as

$$\Psi(\mathbf{r},\xi)=\chi_o(\xi)u_o(\mathbf{r})+\sum a^{(n)}\Phi_n(\mathbf{r},\xi), \tag{80}$$

and this equation holds for all values of r since the boundary conditions on the Φ_n are taken to be outgoing waves at infinity in case the particle can escape by a channel other than the incident one. In this case, the Φ's form a continuum, and the sum in Eq. (80) is to be understood in a generalized sense to include both a sum over discrete states and an integral over the continuum states. Since the boundary conditions are imposed at infinity, introduction of a channel radius is not necessary.

The $a^{(n)}$ are determined by requiring Ψ to satisfy the Schrödinger equation $H\Psi=E\Psi$, and one finds

$$a^{(n)}=\frac{\int\Phi_n(\mathbf{r},\xi)V(\mathbf{r},\xi)\chi_0(\xi)u_0(\mathbf{r})d^3\xi d^3r}{E-\mathcal{E}_n}, \tag{81}$$

where we have used Feshbach's notation in denoting the eigenvalues of Φ_n by $\mathcal{E}_n$. Our $a^{(n)}$ are the same as his Λ_n. One also finds, by multiplying the Schrödinger equation on the left by $\chi_o(\xi)$ and integrating over $d^3\xi$, that

$$(T-\bar{V}-E)u_0(\mathbf{r})+\sum a^{(n)}\int\chi_o(\xi)V(\mathbf{r},\xi)\Phi_n d^3\xi=0. \tag{82}$$

By substituting the $a^{(n)}$ from Eq. (81), one finds that $u_o(\mathbf{r})$ obeys the equation

$$(T+\mathcal{U}-E)u_o(\mathbf{r})=0, \tag{83}$$

where $\mathcal{U}$ is Feshbach's generalized optical potential

$$\mathcal{U}=\bar{V}(r)+\sum_n\frac{\int\chi_0(\xi)V(\mathbf{r},\xi)\Phi_n(\mathbf{r},\xi)d^3\xi\int\Phi_n(\mathbf{r}',\xi)V(\mathbf{r}',\xi)\chi_0(\xi)d^3\xi}{E-\mathcal{E}_n}, \tag{83.1}$$

and $\mathcal{U}$ is an integral (nonlocal) operator. One can easily make further connections between our formalism and his by noting that his matrix H_{ij} is just

$$H_{ij}=\int\chi_i(\xi)H\chi_j(\xi)d^3\xi.$$

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

In the very low energy region in which only elastic scattering is possible, the main difference between the two formalisms results from the fact that the incident channel occurs in our $\Phi^{(p)}$ but not in Feshbach's Φ_n. The width corresponding to this channel is contained in our complex W_p, whereas one must essentially invert a matrix—which can be done in a straightforward fashion—to introduce the widths into the denominators of the scattering amplitude in the formalism of Feshbach. The discussion following Eq. (51) shows that the eigenfunctions $\Phi^{(p)}$ and Φ_n are, aside from the above difference, very nearly the same.

The difference between the formalisms is that one can start from a formalism without joining radius, as Feshbach does, in which case the expression for the wave function is simple, but the transition to the scattering amplitude is somewhat complicated in detail. On the other hand, one can begin with a joining radius, in which case the scattering amplitude is simply obtained, and then carry out various sums as we have done to obtain results in which this radius does not appear. Whereas the former procedure has conceptual advantages, the intermediate formulas seem simpler in the latter procedure and the arguments about the spectrum of the compound states, which we have made to establish perturbation theory, appear to be simpler.

A more general formalism for nuclear reactions has been given by Bloch,[36] and we have referred to his work at several points. By appropriate choice of representation, one can obtain either the Wigner-Eisenbud or Kapur-Peierls theory from his formalism. Making the arguments of this article in his formalism would be completely equivalent to the treatment here, and we have chosen the simpler notation of the less general treatment.

Finally, many of the arguments of this article were first formulated in the Wigner-Eisenbud formalism. This is especially true of the development of Sec. III, 3 as shown by a given reference. However, for uniformity we have chosen to restate most of these arguments in our formalism which is just as suitable for them and is probably simpler for other considerations as discussed in Sec. I.

ACKNOWLEDGMENTS

The author would like to acknowledge his debt to Professor R. E. Peierls for countless helpful suggestions and criticisms relevant to the development presented here. He is also greatly indebted to Dr. C. T. De Dominicis. Many other colleagues at Birmingham have been helpful, and Dr. A. M. Lane has contributed numerous helpful suggestions.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

Bremsstrahlung Cross-Section Formulas and Related Data*

H. W. Koch and J. W. Motz

National Bureau of Standards, Washington, D. C.

I. Introduction 920
II. Bremsstrahlung Cross Sections 921
A. Symbols, Constants, and Energy-Momentum Relations 921
B. Types of Cross-Section Calculations ... 921
C. Bremsstrahlung Cross-Section Formulas and Classification Diagrams 922
(1) Born-Approximation Cross-Section Formulas 926
(2) Extreme-Relativistic Cross-Section Formulas with the Coulomb Correction 927
D. Graphical Representation of the Formulas 927
(1) Dependence of the Bremsstrahlung Spectrum on Electron Energy 929
(2) Dependence of the Bremsstrahlung Spectrum on Photon Angle 929
(3) Screening Effects and Coulomb Corrections 931
E. Corrections for the Cross-Section Formulas 931
(1) Coulomb Corrections 931
(a) Nonrelativistic energies ... 931
(b) Intermediate energies 931
(c) Extreme-relativistic energies 932
(2) High-Frequency-Limit Corrections 933
(3) Screening Corrections 934
F. Comparison of Theory and Experiment 941
(1) Cross-Section Differential in Photon Energy and Angle 942
(2) Cross-Section Differential in Photon Energy 942
(3) Total Cross Section 943
G. Summary 946
III. Electron-Electron Bremsstrahlung 947
A. Maximum Photon Energy 947
B. Cross-Section Formulas for Free Electrons 947
(1) Nonrelativistic Energies 947
(2) Extreme-Relativistic Energies ... 948
C. Cross-Section Formulas with Binding Corrections 948
IV. Thick-Target Bremsstrahlung Production 949
A. Thick-Target Bremsstrahlung Angular Distributions 949
(1) Nonrelativistic and Intermediate Energies 949
(2) Relativistic Energies 950
B. Thick-Target Bremsstrahlung Spectra .. 951
(1) Nonrelativistic and Intermediate Energies 951
(2) Relativistic Energies 952
C. Efficiency for Bremsstrahlung Production 953
(1) Nonrelativistic and Intermediate Energies 955
(2) Relativistic Energies 955
(a) Intermediate thickness targets 955
(b) Thick targets 955
V. Acknowledgments 955

I. INTRODUCTION

CONSIDERABLE information about the bremsstrahlung process has accumulated during the past several years. This information includes various cross-section calculations and measurements, which have helped to provide a more accurate description of the process. Unfortunately this material has never been assembled and integrated in an easily referenced form, although some general reviews[1] on the subject are available. This paper provides a coherent summary of the bremsstrahlung cross-section formulas and related data. The theoretical formulas and their specific limitations are presented in a form convenient for practical calculations. Estimates of their accuracy are given for cases where comparisons can be made with experimental results. Correction factors are indicated in either numerical or analytical form. A brief summary of other data pertaining to electron-electron and to thick target bremsstrahlung is also included. No results are presented for electron and photon polarization effects.

Section IIB briefly discusses the problem of making exact cross-section calculations and indicates the general types of calculations that have been completed. A summary of the various cross-section formulas is given in Sec. IIC. Section IID gives useful graphical information derived from the various formulas in IIC. Section IIE lists corrections that can be applied to the above

* The research reported here was supported jointly by the U. S. Atomic Energy Commission, and by the U. S. Air Force, through the Office of Scientific Research of the Air Research and Development Command.

[1] H. Bethe and E. Salpeter, *Encyclopedia of Physics* (Springer-Verlag, Berlin, 1957), Vol. 35, p. 425; S. T. Stephenson, *ibid.*, Vol. 30, p. 337.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

formulas. In Sec. IIF, experimental bremsstrahlung cross sections are compared with the theoretical results contained in Secs. IIC and D. Conclusions with regard to the accuracy of the theory are presented in Sec. IIG. Section III summarizes the very sparse material available on electron-electron bremsstrahlung. Finally, Sec. IV gives a brief treatment of thick target bremsstrahlung with information on the bremsstrahlung angular distributions (IVA), the spectra (IVB), and the production efficiencies (IVC).

II. BREMSSTRAHLUNG CROSS SECTIONS

The cross sections discussed in this section apply to the bremsstrahlung process[2] in which an electron is decelerated in the field of an atomic nucleus. These cross sections give direct estimates of the properties of the radiation emitted when electrons are incident on *thin*[3] targets, and provide basic data for analyzing the *thick* target bremsstrahlung considered in Sec. IV.

A. Symbols, Constants, and Energy-Momentum Relations

E_0, E = initial and final total energy of the electron in a collision, in m_0c^2 units.†

$\mathbf{p}_0, \mathbf{p}$ = initial and final momentum of the electron in a collision, in m_0c units.

T_0, T = initial and final kinetic energy of the electron in a collision, in m_0c^2 units.†

$k, \mathbf{k}$ = energy and momentum of the emitted photon, in m_0c^2 and m_0c units.†

$\mathcal{E}_0$ = total energy of an electron incident on a thick target, in m_0c^2 units.†

$\mathcal{T}_0$ = kinetic energy of an electron incident on a thick target, in m_0c^2 units.†

θ_0, θ = angles of $\mathbf{p}_0$ and $\mathbf{p}$ with respect to $\mathbf{k}$.

ϕ = angle between the planes $(\mathbf{p}_0,\mathbf{k})$ and $(\mathbf{p},\mathbf{k})$.

$d\Omega_k$ = element of solid angle, $\sin\theta_0 d\theta_0 d\phi$, in the direction of $\mathbf{k}$.

$d\Omega_p$ = element of solid angle, $\sin\theta d\theta d\phi$, in the direction of $\mathbf{p}$.

$\mathbf{q}$ = momentum transferred to the nucleus, in m_0c units.
$= \mathbf{p}_0 - \mathbf{p} - \mathbf{k}$; $q^2 = p_0^2 + p^2 + k^2 - 2p_0k\cos\theta_0 + 2pk\cos\theta - 2p_0p(\cos\theta\cos\theta_0 + \sin\theta\sin\theta_0\cos\phi)$.

β_0, β = ratio of the initial and final electron velocity in a collision to the velocity of light.

Z = atomic number of target material.

$d\sigma_j$ = bremsstrahlung cross section, differential with respect to the parameter j, in units of cm² per atom per incident electron.

$d\tau$ = volume element.

$\mathbf{r}$ = radius vector from a center, in units of the Compton wavelength, $\lambda\!\!\!^{-}_0$.

α = angle of $\mathbf{k}$ with respect to the direction of the electron beam incident on a thick target.

N_0 = 6.03×10^{23} atoms (or molecules) per mole.

c = 3.00×10^{10} cm per sec.

e = 4.80×10^{-10} esu = 1.60×10^{-19} coulomb.

e^2 = 1.44×10^{-13} Mev cm.

h = $h/2\pi = 6.58\times10^{-22}$ Mev sec = 1.05×10^{-27} erg sec.

hc = 12.4 kev-angstroms.

$\hbar c$ = 1.97×10^{-11} Mev cm.

hc/e^2 = 137.

m_0 = 9.11×10^{-28} g (electron mass).

m_0c^2 = 0.511 Mev.

$\lambda\!\!\!^{-}_0$ = $\hbar/m_0c = 3.86\times10^{-11}$ cm (Compton wavelength).

r_0 = $e^2/m_0c^2 = \lambda\!\!\!^{-}_0/137 = 2.82\times10^{-13}$ cm (classical electron radius).

$\bar{\phi}$ = $Z^2r_0^2/137 = Z^2\ 5.78\times10^{-28}$ cm².

a_0 = $\hbar^2/m_0e^2 = 0.530\times10^{-8}$ cm (radius of hydrogen atom).
$= 137\lambda\!\!\!^{-}_0 = (137)^2 r_0$

ϕ_0 = $8\pi r_0^2/3$ (Thomson formula) = 6.64×10^{-25} cm².

e^2/a_0 = $2I_0 = 27.2$ ev.

I_0 = ionization energy of hydrogen atom = $1/2(137)^2$ in m_0c^2 units.

Z^2I_0 = ionization energy of K electron ($if \ll 1$).

1 Mev = 1.60×10^{-6} erg.

$E_0^2 = p_0^2 + 1$, $E^2 = p^2 + 1$.

$E_0 = T_0 + 1$, $E = T + 1$.

$$E_0 = \frac{1}{(1-\beta_0^2)^{\frac{1}{2}}}, \quad E = \frac{1}{(1-\beta^2)^{\frac{1}{2}}}$$

$E_0 = k + E$.

$p_0 = [T_0(T_0+2)]^{\frac{1}{2}}$, $p = [T(T+2)]^{\frac{1}{2}}$

$$p_0 = \frac{\beta_0}{(1-\beta_0^2)^{\frac{1}{2}}}, \quad p = \frac{\beta}{(1-\beta^2)^{\frac{1}{2}}}.$$

$$\beta_0 = \frac{p_0}{E_0}, \quad \beta = \frac{p}{E}.$$

B. Types of Cross-Section Calculations

The bremsstrahlung cross section $d\sigma$, for single photon emission in a large cubic box of side L, is given by the transition probability per atom per electron divided by the incoming electron velocity. This cross section

[2] Except for electron-electron bremsstrahlung which is briefly considered in Sec. III, no results are presented for other bremsstrahlung processes, involving for example the acceleration of positrons or protons.

[3] A target is defined to be thin if both the electron scattering and energy loss processes have a negligible influence on the energy and angular distributions of the bremsstrahlung. Order of magnitude estimates of such thin targets for particular energy regions can be found in the references listed in Sec. IIG.

† This system of units for the symbols is used consistently throughout this paper. For cases in which the data are given in Mev units, these symbols have the multiplicative factor 0.51; for example, the kinetic energy in Mev units is represented by the quantity $0.51T_0$.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

can be expressed in dimensions of cm² as

$$d\sigma=\frac{w}{(p_0c/E_0)}\left(\frac{\hbar}{m_0c}\right)^3 L^3, \qquad \text{(II-1)}$$

where

$$w=(2\pi/\hbar)\rho_f|H_{if}|^2. \qquad \text{(II-2)}$$

The term ρ_f is the density of final states and can be written as

$$\rho_f=\frac{pEk^2dkd\Omega_kd\Omega_pL^6}{(2\pi)^6m_0c^2}. \qquad \text{(II-3)}$$

The term H_{if} is the matrix element for the transition of the system from an initial state before the emission of the photon to a final state after the emission. The quantity $|H_{if}|^2$ in formula (II-2) can be written as

$$|H_{if}|^2=\left(\frac{2\pi}{k}\frac{e^2}{\hbar c}\right)(m_0c^2)^2 \times\left|\int\psi_f^*(\lambda^*\cdot\alpha)e^{-i\mathbf{k}\cdot\mathbf{r}}\psi_i d\tau\right|^2 L^{-9}. \qquad \text{(II-4)}$$

In the foregoing, λ is the unit polarization vector of the photon, α is the Dirac matrix, and ψ_i and ψ_f are the Dirac wave functions for the initial and final electrons, respectively. Therefore the cross section in cm² can be written as

$$d\sigma=\frac{137r_0^2}{(2\pi)^4}\frac{pE_0E}{p_0}\left|\int\psi_f^*(\lambda^*\cdot\alpha)e^{-i\mathbf{k}\cdot\mathbf{r}}\psi_i d\tau\right|^2 \times kdkd\Omega_kd\Omega_p. \qquad \text{(II-5)}$$

The important quantity to be evaluated is the matrix element H_{if}, defined in formula (II-4).

The problem of evaluating an "exact" expression for the cross section involves, therefore, the use in the matrix element of "exact" wave functions, which describe an electron in a screened, nuclear Coulomb field. It is not possible to solve the Dirac wave equation in closed form for an electron in a Coulomb field, primarily because the wave function must be represented as an infinite series.[4] Therefore, various approximate wave functions and procedures have been used.

The cross-section calculations that have been made may be classified either as nonrelativistic or relativistic depending on whether the Schrödinger or Dirac form of the Hamiltonian is used for the electron and field system. The calculations have been carried out with (a) nonrelativistic Coulomb wave functions (Sommerfeld); (b) relativistic Coulomb wave functions (Sommerfeld-Maue) valid to first order in $(Z/137)^2/l$, where l is the angular momentum quantum number that is the summation index in the expansion of the wave function; and (c) free-particle wave functions perturbed to first order in Z (Born-approximation procedure).

The nonrelativistic cross-section formulas derived in the dipole approximation by Sommerfeld[5] with Coulomb wave functions have a complicated form with hypergeometric functions and are difficult to evaluate. Some numerical estimates[6] of the Sommerfeld cross sections have been made for selected values of the electron energy, the target atomic number, and the photon energy. However, the theory is only valid when β_0 is small compared to unity, and can be expected to break down for initial electron energies greater than a few kilovolts. In addition, the theory disregards screening effects, which are important for very low energies and for targets with high atomic numbers. Because of these limitations, results of the Sommerfeld theory are not presented here.

Cross-section calculations with relativistic Coulomb wave functions (Sommerfeld-Maue) including screening corrections have been made by Olsen, Maximon, and Wergeland[7] and by Olsen and Maximon.[8] Their formulas are valid only in the extreme-relativistic region (above 50 Mev). Their results have the form of an additive correction factor to the Born-approximation formulas.

The cross-section formulas calculated by the Born-approximation procedure with free-particle wave functions are available in a relatively simple analytical form for nonrelativistic and relativistic energies, with or without screening. In general, the Born approximation theory becomes less reliable as (a) the atomic number of the target increases, (b) the initial electron energy decreases, and (c) the photon energy approaches the high-frequency limit. In spite of their limited validity, the Born-approximation formulas have been surprisingly successful in predicting the properties of the bremsstrahlung radiation. Even where there is a breakdown of the Born approximation, the accuracy of the cross-section formulas is still reasonably good, and in the worst cases (except at the high-frequency limit), they can be expected to give at least the correct order of magnitude. Therefore, this paper emphasizes the Born-approximation cross-section formulas and includes various theoretical and empirical corrections to these formulas. Detailed references to the many papers in which these formulas are derived are given in Table III.

C. Bremsstrahlung Cross-Section Formulas and Classification Diagrams

A general classification of the various differential forms of the bremsstrahlung cross section is presented in Chart 1 for the Born-approximation formulas and in Chart 2 for the extreme-relativistic formulas that con-

[4] A detailed discussion of this problem is given by H. A. Bethe and L. C. Maximon, Phys. Rev. 93, 768 (1954).

[5] A. Sommerfeld, *Wellenmechanik* (Frederick Ungar, New York, 1950), Chap. 7.

[6] P. Kirkpatrick and L. Wiedmann, Phys. Rev. 67, 321 (1945).

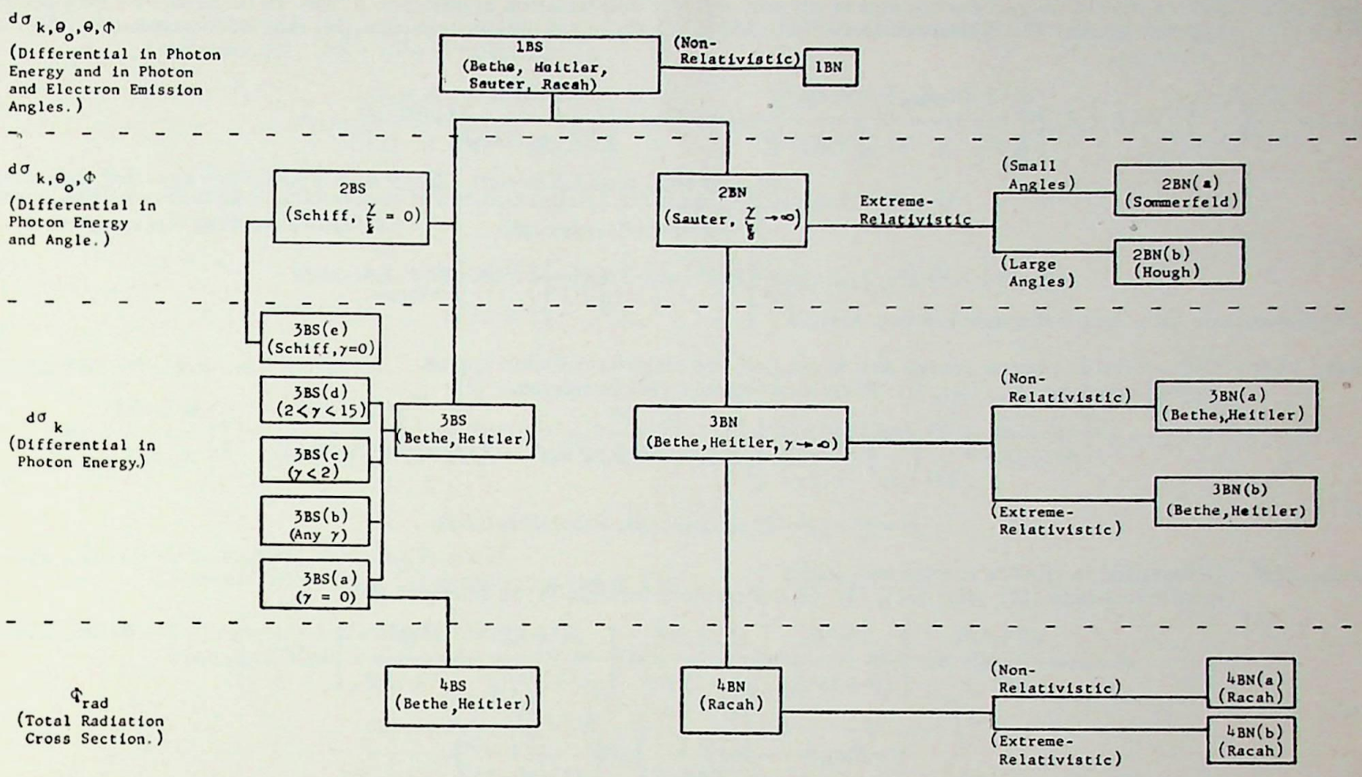
[7] Olsen, Maximon, and Wergeland, Phys. Rev. 106, 27 (1957).

[8] H. Olsen and L. C. Maximon, Phys. Rev. 114, 887 (1959).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

CHART 1

Born Approximation Cross Section Formulas



tain the Coulomb correction. The formulas represented are summed over the directions of the electron spin and the photon polarization vectors, and thus do not include all of the possible differential forms of the cross section. The primary formula gives the cross section that is differential in photon energy and in photon and electron emission angles. The remaining formulas that branch out from this starting point are divided into two main groups that are designated as screened or non-screened. Further subdivisions are made; these depend on the type of screening approximation, and on whether nonrelativistic, extreme-relativistic, small-angle, or large-angle approximations are used. For most of the cases, the charts include the names of the principal authors associated with a particular formula.

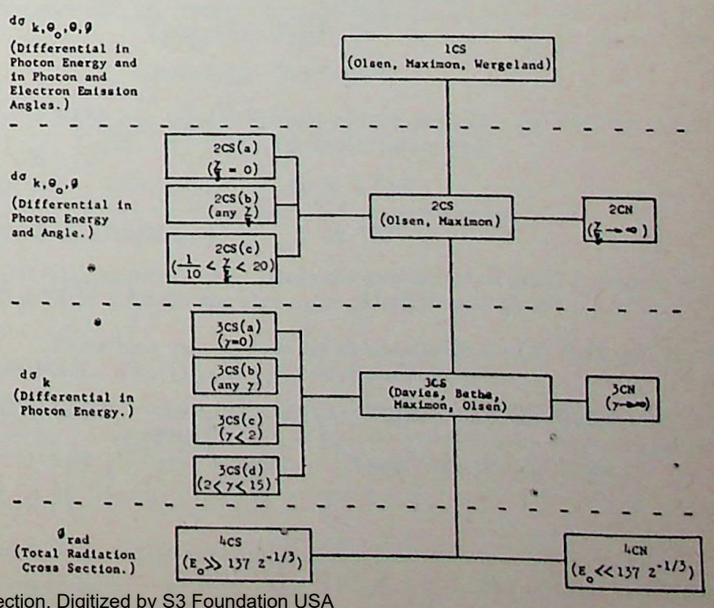
The formulas are identified as follows: (a) the number applies to a particular differential form of the cross section; (b) the first letter indicates either B for Born approximation (Chart 1) or C for Coulomb correction (Chart 2); (c) the second letter indicates either S for screening or N for no screening; and (d) the last letter a, b, or c indicates further subdivisions for specific approximations. The following notation has been adopted here. The differential forms of the bremsstrahlung cross section are designated by the symbol, $d\sigma_{\alpha,\beta,\ldots}$. This symbol is the bremsstrahlung cross section that is differential only with respect to the parameters given by the subscripts $\alpha, \beta, \cdots$, and is explicitly defined by $d\sigma_{\alpha,\beta\ldots} \equiv (d^n\sigma/d\alpha d\beta\cdots)d\alpha d\beta\cdots$. The unit of the cross section, $d\sigma_{\alpha,\beta\ldots}$, is cm^2 per atom per incident electron.

The symbols and definitions for the specific cross sections are as follows.

(a) $d\sigma_{k,\theta_0,\theta,\phi}$ is the bremsstrahlung cross section that is differential with respect to the photon energy, k, and to the photon and electron emission angles, θ_0, θ, and ϕ. This formula contains the parameters E_0, Z, k, θ_0, θ, and ϕ.

CHART 2

Extreme-Relativistic Cross Section Formulas with Coulomb Correction



CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

TABLE I. Born-approximation cross-section formulas.

Formula 1BS—Differential in photon energy and in photon and electron emission angles.
Approximation (H). Reference formulas: (13) in reference (a), (2) in reference (b), (13) in reference (c), (29) in reference (e).

$$d\sigma_{k,\theta_0,\theta,\phi}=\frac{Z^2}{137}\left(\frac{r_0}{2\pi}\right)^2[1-F(q,Z)]^2\frac{dk}{k}\frac{p}{p_0}\frac{d\Omega_k d\Omega_p}{q^4}\left\{\frac{p^2\sin^2\theta}{(E-p\cos\theta)^2}(4E_0^2-q^2)+\frac{p_0^2\sin^2\theta_0}{(E_0-p_0\cos\theta_0)^2}(4E^2-q^2)\right.$$
$$\left.-\frac{2pp_0\sin\theta\sin\theta_0\cos\phi\,(4EE_0-q^2)}{(E-p\cos\theta)(E_0-p_0\cos\theta_0)}+\frac{2k^2(p^2\sin^2\theta+p_0^2\sin^2\theta_0-2pp_0\sin\theta\sin\theta_0\cos\phi)}{(E-p\cos\theta)(E_0-p_0\cos\theta_0)}\right\},$$

where

$$q^2=p^2+p_0^2+k^2-2p_0k\cos\theta_0+2pk\cos\theta-2p_0p(\cos\theta\cos\theta_0+\sin\theta\sin\theta_0\cos\phi)$$

and $F(q,z)$ = atomic form factor discussed in Sec. IIE(3).

Formula 1BN—Differential in photon energy and in photon and electron emission angles.
Approximations (H), (B), (I). Reference formula (17) in reference (c).

$$d\sigma_{k,\theta_0,\theta,\phi}=\frac{Z^2}{137}\left(\frac{r_0}{\pi}\right)^2\frac{dk}{k}\frac{p}{p_0}\frac{d\Omega_k d\Omega_p}{q^4}\{p^2\sin^2\theta+p_0^2\sin^2\theta_0-2pp_0\sin\theta\sin\theta_0\cos\phi\},$$

where

$$q^2=p^2+p_0^2-2pp_0(\cos\theta\cos\theta_0+\sin\theta\sin\theta_0\cos\phi).$$

Formula 2BS—Differential in photon energy and angle.
Approximations (H), (G), (M), (J), (K). Reference formula (1) in reference (h).

$$d\sigma_{k,\theta_0}=\frac{4Z^2r_0^2}{137}\frac{dk}{k}y\,dy\left\{\frac{16y^2E}{(y^2+1)^4E_0}-\frac{(E_0+E)^2}{(y^2+1)^2E_0^2}+\left[\frac{E_0^2+E^2}{(y^2+1)^2E_0^2}-\frac{4y^2E}{(y^2+1)^4E_0}\right]\ln M(y)\right\},$$

where

$$y=E_0\theta_0,\quad\frac{1}{M(y)}=\left(\frac{k}{2E_0E}\right)^2+\left(\frac{Z^{\frac{1}{3}}}{111(y^2+1)}\right)^2.$$

Comment: This formula becomes Formula 2BN(a) when $Z=0$ in $M(y)$.

Formula 2BN—Differential in photon energy and angle.
Approximations (H), (B). Reference formulas: (11) in reference (d), (4.1) in reference (f).

$$d\sigma_{k,\theta_0,\phi}=\frac{Z^2r_0^2}{8\pi137}\frac{dk}{k}\frac{p}{p_0}d\Omega_k\left\{\frac{8\sin^2\theta_0(2E_0^2+1)}{p_0^2\Delta_0^4}-\frac{2(5E_0^2+2EE_0+3)}{p_0^2\Delta_0^2}-\frac{2(p_0^2-k^2)}{Q^2\Delta_0^2}+\frac{4E}{p_0^2\Delta_0}\right.$$
$$+\frac{L}{pp_0}\left[\frac{4E_0\sin^2\theta_0(3k-p_0^2E)}{p_0^2\Delta_0^4}+\frac{4E_0^2(E_0^2+E^2)}{p_0^2\Delta_0^2}+\frac{2-2(7E_0^2-3EE_0+E^2)}{p_0^2\Delta_0^2}+\frac{2k(E_0^2+EE_0-1)}{p_0^2\Delta_0}\right]$$
$$\left.-\left(\frac{4\epsilon}{p\Delta_0}\right)+\left(\frac{\epsilon^Q}{pQ}\right)\left[\frac{4}{\Delta_0^2}-\frac{6k}{\Delta_0}-\frac{2k(p_0^2-k^2)}{Q^2\Delta_0}\right]\right\},$$

where

$$L=\ln\left[\frac{EE_0-1+pp_0}{EE_0-1-pp_0}\right];\quad\Delta_0=E_0-p_0\cos\theta_0;\quad\epsilon=\ln\left[\frac{E+p}{E-p}\right];\quad\epsilon^Q=\ln\left[\frac{Q+p}{Q-p}\right]$$
$$Q^2=p_0^2+k^2-2p_0k\cos\theta_0.$$

Formula 2BN(a)—Differential in photon energy and angle.
Approximations (H), (B), (J), (K).

$$d\sigma_{k,\theta_0,\phi}=\frac{2Z^2r_0^2}{\pi137}\frac{E}{E_0}\frac{dk}{k}d\Omega_k\left\{\frac{16(\theta_0E_0)^2E_0^2}{(1+\theta_0^2E_0^2)^4}-\frac{(E_0+E)^2E_0}{E(1+\theta_0^2E_0^2)^2}+2\ln\left(\frac{EE_0}{k}\right)\left[\frac{(E^2+E_0^2)E_0}{E(1+\theta_0^2E_0^2)^2}-\frac{4\theta_0^2E_0^4}{(1+\theta_0^2E_0^2)^4}\right]\right\}.$$

Comment: This formula was obtained from formula 2BN by making the high-energy and small-angle approximations. The same result is obtainable from formula 2BS by setting $Z=0$ in $M(y)$.

Formula 2BN(b)—Differential in photon energy and angle.
Approximations (H), (B), (J), (L). Reference formula (8) in reference (i).

$$d\sigma_{k,\theta_0,\phi}=\frac{Z^2r_0^2}{4\pi137}\frac{E}{E_0^3}\frac{dk}{k}\frac{d\Omega_k}{(1-\cos\theta_0)^2}\left\{\left(\frac{E_0^2+E^2}{EE_0}\right)\sin^2\theta_0\ln\frac{2E_0E}{k}-\frac{(5E_0+2E)}{E_0}+\frac{2[E^2-2E_0^2\ln(k/E_0)]}{E_0E}[1-\cos\theta_0]\right.$$
$$\left.-\frac{E[Q^2+E(k-E_0\cos\theta_0)]}{E_0Q^2(1-\cos\theta_0)}-\frac{E_0k(1-\cos\theta_0)}{EQ^3}[3Q^2+E(E_0+k)]\ln\left(\frac{Q+E}{Q-E}\right)\right\}.$$

where $Q^2=E^2+2kE_0(1-\cos\theta_0)$.

CC-0. Gurukul Kangri University, Haridwar Collection. Digitized by S3 Foundation USA

TABLE I.—(*Continued*).

Formula 3BS—Differential in photon energy.
Approximations (H), (J). Reference formulas: (31), (34), (35) in reference (a); (62) in reference (b); (21), (26) in reference (c); (56)–(58b) in reference (j).

$$d\sigma_k=\frac{4Z^2r_0^2}{137}\frac{dk}{k}\left\{\left(1+\left(\frac{E}{E_0}\right)^2\right)\left[\frac{\phi_1(\gamma)}{4}-\tfrac{1}{3}\ln Z\right]-\frac{2}{3}\frac{E}{E_0}\left[\frac{\phi_2(\gamma)}{4}-\tfrac{1}{3}\ln Z\right]\right\}.$$

Formula 3BS(a)—Complete screening ($\gamma=0$ or $\phi_1(\gamma=0)=4\ln 183$; $\phi_2(\gamma=0)=\phi_1(\gamma=0)-\frac{2}{3}$).
Formula 3BS with $\gamma=0$.

$$d\sigma_k=\frac{4Z^2r_0^2}{137}\frac{dk}{k}\left\{\left[1+\left(\frac{E}{E_0}\right)^2-\frac{2}{3}\frac{E}{E_0}\right]\ln(183Z^{-\frac{1}{3}})+\frac{1}{9}\frac{E}{E_0}\right\}.$$

Formula 3BS(b)—Arbitrary screening.

$$d\sigma_k=\frac{4Z^2r_0^2}{137}\frac{dk}{k}\left\{\left(1+\left(\frac{E}{E_0}\right)^2\right)\left[\int_\delta^1(q-\delta)^2(1-F(q))^2\frac{dq}{q^3}+1\right]-\frac{2}{3}\frac{E}{E_0}\left[\int_\delta^1\left(q^3-6\delta^2q\ln\frac{q}{\delta}+3\delta^2q-4\delta^3\right)(1-F(q))^2\frac{dq}{q^4}+\frac{5}{6}\right]\right\},$$

where

$$\delta=k/(2E_0E).$$

Formula 3BS(c)—Intermediate screening I ($\gamma<2$).
Formula 3BS with $\phi_1(\gamma)$ and $\phi_2(\gamma)$ given in Fig. 1.

Formula 3BS(d)—Intermediate screening II ($2<\gamma<15$).
Formula 3BS with $\phi_1(\gamma)=\phi_2(\gamma)=19.19-4\ln\gamma-4c(\gamma)$ with $c(\gamma)$ given in Fig. 2.

$$d\sigma_k=\frac{4Z^2r_0^2}{137}\frac{dk}{k}\left[1+\left(\frac{E}{E_0}\right)^2-\frac{2}{3}\frac{E}{E_0}\right]\left[\ln\frac{2E_0E}{k}-\tfrac{1}{2}-c(\gamma)\right].$$

Formula 3BS(e)—Differential in photon energy.
Approximations (H), (G), (M), (J). Reference formula (3) in reference (h).

$$d\sigma_k=\frac{2Z^2r_0^2}{137}\frac{dk}{k}\left\{\left(1+\left(\frac{E}{E_0}\right)^2-\frac{2}{3}\frac{E}{E_0}\right)\left(\ln M(0)+1-\frac{2}{b}\tan^{-1}b\right)+\frac{E}{E_0}\left[\frac{2}{b^2}\ln(1+b^2)+\frac{4(2-b^2)}{3b^2}\tan^{-1}b-\frac{8}{3b^2}+\frac{2}{9}\right]\right\},$$

where

$$b=\left(\frac{2E_0EZ^{\frac{1}{3}}}{111k}\right);\quad\frac{1}{M(0)}=\left(\frac{k}{2E_0E}\right)^2+\left(\frac{Z^{\frac{1}{3}}}{111}\right)^2.$$

Comment: This formula is obtained from formula 50 of reference (b).

Formula 3BN—Differential in photon energy.
Approximations (H), (B). Reference formulas: (15) in reference (a), (16) in reference (c), (17) in reference (d), (37) in reference (e).

$$d\sigma_k=\frac{Z^2r_0^2}{137}\frac{dk}{k}\frac{p}{p_0}\left\{\tfrac{4}{3}-2E_0E\left(\frac{p^2+p_0^2}{p^2p_0^2}\right)+\frac{\epsilon_0E}{p_0^3}+\frac{\epsilon E_0}{p^3}-\frac{\epsilon\epsilon_0}{p_0p}+L\left[\frac{8E_0E}{3p_0p}+\frac{k^2(E_0^2E^2+p_0^2p^2)}{p_0^3p^3}+\frac{k}{2p_0p}\left(\left(\frac{E_0E+p_0^2}{p_0^3}\right)\epsilon_0-\left(\frac{E_0E+p^2}{p^3}\right)\epsilon+\frac{2kE_0E}{p^2p_0^2}\right)\right]\right\},$$

where

$$L=2\ln\left[\frac{E_0E+p_0p-1}{k}\right];\quad\epsilon_0=\ln\left(\frac{E_0+p_0}{E_0-p_0}\right);\quad\epsilon=\ln\left(\frac{E+p}{E-p}\right).$$

Formula 3BN(a)—Differential in photon energy.
Approximations (H), (B), (I). Reference formula (18) in reference (c).

$$d\sigma_k=\frac{Z^2r_0^2}{137}\frac{16}{3}\frac{dk}{k}\frac{1}{p_0^2}\ln\left(\frac{p_0+p}{p_0-p}\right).$$

Formula 3BN(b)—Differential in photon energy.
Approximations (H), (B), (J). Reference formulas: (16) in reference (a), (21) in reference (c), (56) in reference (j).

$$d\sigma_k=\frac{4Z^2r_0^2}{137}\frac{dk}{k}\left[1+\left(\frac{E}{E_0}\right)^2-\frac{2}{3}\frac{E}{E_0}\right]\left[\ln\left(\frac{2E_0E}{k}\right)-\tfrac{1}{2}\right].$$

Comment: This formula results directly from Formula 3BS(e) when $b\to 0$ and from formula 3BS when $\gamma\to\infty$.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

TABLE I.—*(Continued)*.

Formula 4BS—Total radiation cross section.
Approximations (H), (F), (J). Reference formulas: (47) in reference (a), (34) in reference (c), (62) in reference (j).

$$\phi_{\text{rad}}=\frac{4Z^2r_0^2}{137}\left[\ln(183Z^{-\frac{1}{3}})+\frac{1}{18}\right].$$

Formula 4BN—Total radiation cross section.
Approximations (H), (C). Reference formulas: (29) in reference (c), (41) in reference (e).

$$\phi_{\text{rad}}=\frac{Z^2r_0^2}{137}\left\{\frac{(12E_0^2+4)}{3E_0p_0}\ln(E_0+p_0)-\frac{(8E_0+6p_0)}{3E_0p_0^2}[\ln(E_0+p_0)]^2-\tfrac{4}{3}+\frac{2}{E_0p_0}[F(x)]\right\},$$

where

$$F(x)=\int_0^x\frac{\ln(1+y)}{y}dy \quad \text{and} \quad x=2p_0(E_0+p_0).$$

For small x, F can be expanded in the power series:

$$F(x)=x-\frac{x^2}{4}+\frac{x^3}{9}-\frac{x^4}{16}+\cdots.$$

For large x, F is given by

$$F(x)=\tfrac{1}{6}\pi^2+\tfrac{1}{2}(\ln x)^2-F(1/x).$$

Formula 4BN(a)—Total radiation cross section.
Approximations (H), (C), (I). Reference formulas: (32) in reference (c), (21) in reference (d).

$$\phi_{\text{rad}}=(16/3)(Z^2r_0^2/137).$$

Formula 4BN(b)—Total radiation cross section.
Approximations (H), (C), (J). Reference formulas: (33) in reference (c), (41′) in reference (e), (61) in reference (j), (46) in (a), (22) in (d).

$$\phi_{\text{rad}}=\frac{4Z^2r_0^2}{137}(\ln 2E_0-\tfrac{1}{3}).$$

(b) $d\sigma_{k,\theta_0,\phi}$ is the bremsstrahlung cross section that is differential with respect to the photon energy, k, and the emission angles θ_0 and ϕ. It can be obtained by integrating the differential cross section in (a) over the direction of the outgoing electron. This formula contains the parameters E_0, Z, k, and θ_0.

(c) $d\sigma_k$ is the bremsstrahlung cross section that is differential with respect to the photon energy k. It can be obtained by integrating the differential cross section in (a) over the emission directions of the photon and the electron. This formula contains the parameters E_0, Z, and k.

(d) ϕ_{rad} is the only cross-section symbol used here that does not represent a differential form of the bremsstrahlung cross section. It is equal to the quantity, $(1/E_0)\int_0^{T_0}kd\sigma_k$. This form of a total bremsstrahlung cross section integrated over photon energy and photon and electron emission angles was introduced by Heitler,[9] who has defined it as the cross section for the energy lost by radiation. This formula contains the parameters E_0 and Z.

The Born-approximation formulas that apply to Chart 1 are presented in Table I, and the extreme-relativistic formulas with the Coulomb correction that apply to Chart 2 are presented in Table II. The important references and approximations for the formulas in Tables I and II are listed in Table III. The explicit expressions for the formulas in Tables I and II are not necessarily the same as the formulas in the original references because the attempt is made to use consistent units and symbols, with energies and momenta expressed in m_0c^2 and m_0c units, respectively.

(1) *Born-Approximation Cross-Section Formulas*

The Born-approximation calculations require the initial and final electron kinetic energies in a collision to be large enough to satisfy the conditions: $(2\pi Z/137\beta_0)$, $(2\pi Z/137\beta)\ll 1$. At extreme-relativistic energies, the cross sections predicted by the Born-approximation formulas are larger than the true cross sections. For example, the value of the total cross section predicted for lead by the Born-approximation formula is about 10% larger than the value predicted by more accurate formulas.[7] At very low energies, the situation is reversed and the Born-approximation cross section is smaller than the true cross section. The energy region in which the Born-approximation formulas require only small corrections is approximately between 4 and 10 Mev for the initial electron kinetic energy. As a rough guide, it is estimated that Born-approximation formulas for the total radiation cross section, ϕ_{rad}, are correct to within 10% for initial electron kinetic energies above 2 Mev and within a factor of two below 2 Mev.

[9] W. Heitler, *The Quantum Theory of Radiation* (Oxford University Press, London, 1954), third edition, p. 242.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

(2) *Extreme-Relativistic Cross-Section Formulas with the Coulomb Correction*

The formulas in Table II are valid for arbitrary Z and have been developed in a series of papers by Bethe and Maximon,[10] Olsen,[11] Olsen, Maximon and Wergeland,[7] Olsen and Maximon,[8,12] and Davies, Bethe, and Maximon.[13] Their calculations were carried out (a) with Sommerfeld-Maue wave functions, and (b) with the extreme relativistic approximations: E_0, E, $k \gg 1$, and $p_0\theta_0 \sim 1$. These formulas are estimated[10] to have an accuracy of the order of $(Z/137)^2(\ln E/E)$ which is better than 2% for electron kinetic energies above 50 Mev and for Z equal to 80.

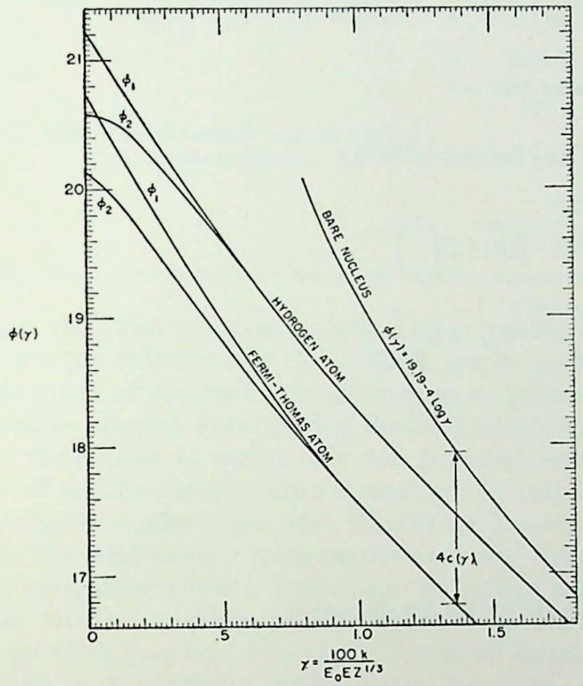


FIG. 1. Screening factors,[28] $\phi_1(\gamma)$ and $\phi_2(\gamma)$, for electron-nuclear bremsstrahlung plotted as a function of $\gamma = 100k/E_0EZ^{\frac{1}{3}}$. The curve marked "Hydrogen atom" was calculated[41] with exact wave functions. The curves for the Thomas-Fermi atom and a bare nucleus differ by the quantity $4c(\gamma)$, where the function $c(\gamma)$ is plotted in Fig. 2.

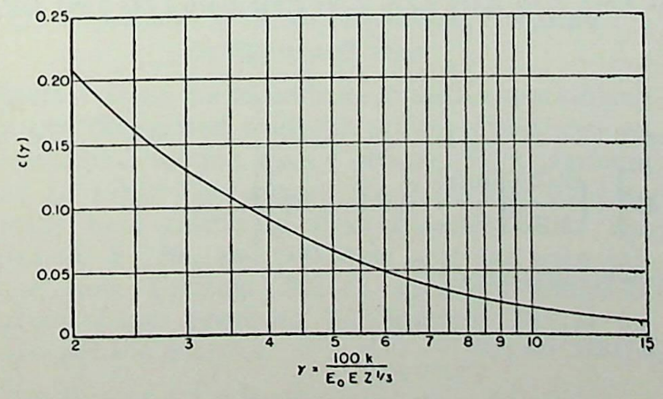


FIG. 2. Screening factor,[28] $c(\gamma)$, for electron-nuclear bremsstrahlung plotted as a function of $\gamma = 100k/E_0EZ^{\frac{1}{3}}$.

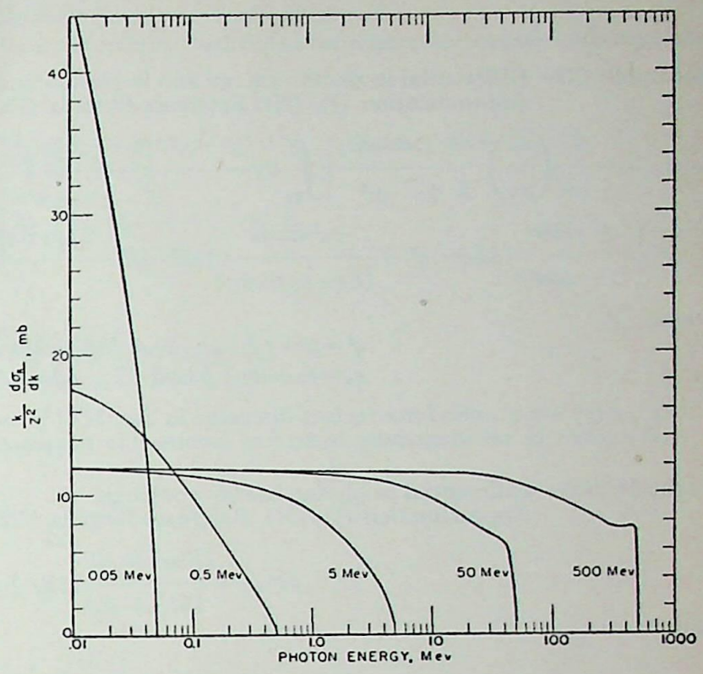


FIG. 3. Dependence of the Born-approximation cross section integrated over the photon directions on the photon and electron energy. The ordinate values for these curves are obtained from Formula 3BN for 0.05- and 0.50-Mev electrons, and from Formula 3BS (e) for 5-, 50-, and 500-Mev electrons.

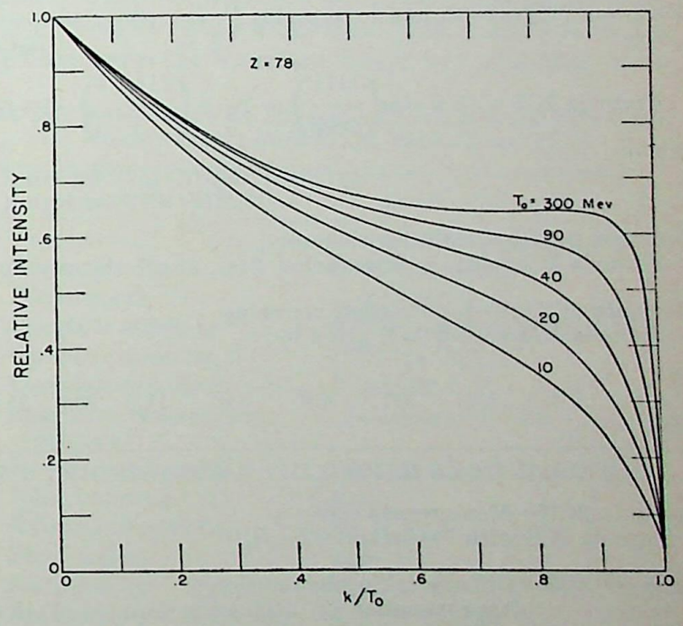


FIG. 4. Dependence of the bremsstrahlung spectrum shape on the electron kinetic energy for a platinum target ($Z=78$). The relative intensity (defined as proportional to the product of the photon energy and number per unit time) is integrated over the photon direction and is normalized to unity for zero photon energies. The intensity values were computed from Formula 3BS(e).

D. Graphical Representations of the Formulas

A general picture of the dependence of the cross-section formulas in Sec. IIC on the electron and photon

[10] H. Bethe and L. C. Maximon, Phys. Rev. **93**, 768 (1954).
[11] H. Olsen, Phys. Rev. **99**, 1335 (1955).
[12] H. Olsen and L. C. Maximon, Phys. Rev. **110**, 589 (1958).
[13] Davies, Bethe, and Maximon, Phys. Rev. **93**, 788 (1954).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

TABLE II. Bremsstrahlung cross-section formulas with Coulomb correction.

Formula 1CS—Differential in photon energy and in photon and electron emission angles.
Approximation (J), (N). Reference formula (7b.13) in reference (m).

$$d\sigma_{k,\theta_0,\theta,\phi}=\frac{Z^2}{137}\left(\frac{r_0}{2\pi}\right)^2\frac{dk}{k}\frac{p}{p_0}\frac{d\Omega_k d\Omega_p}{q_\perp^2}\left|\int_{q_z}^{\infty}dy\frac{(y^2-q_z^2)^{\frac{1}{2}}(1-F(y))}{y}\int_0^{\infty}d\rho\cdot\rho J_1(q_\perp\rho)J_1[\rho(y^2-q_z^2)^{\frac{1}{2}}]\exp\left\{2ia\int_0^{\infty}\frac{J_0(x\rho)[1-F(x)]dx}{x}\right\}\right|^2$$

$$\times\left\{\frac{p^2\sin^2\theta}{(E-\cos\theta)^2}(4E_0^2-q^2)+\frac{p_0^2\sin^2\theta_0}{(E_0-p_0\cos\theta_0)^2}(4E^2-q^2)-\frac{2pp_0\sin\theta\sin\theta_0\cos\phi(4EE_0-q^2)+2k^2(p^2\sin^2\theta+p_0^2\sin^2\theta_0-2pp_0\sin\theta\sin\theta_0\cos\phi)}{(E-p\cos\theta)(E_0-p_0\cos\theta_0)}\right\},$$

where

$$q^2=p^2+p_0^2+k^2-2p_0k\cos\theta_0+2pk\cos\theta-2p_0p(\cos\theta\cos\theta_0+\sin\theta\sin\theta_0\cos\phi)$$
$$q_z=p_0\cos\theta_0+p\cos\theta-k;\quad q_\perp^2=p_0^2\sin^2\theta_0+p^2\sin^2\theta-2p_0p\sin\theta_0\sin\theta\cos\phi.$$

$F(x)$, $F(y)$ are atomic form factors discussed in Sec. IIE(3) and functions of the momentum variables x and y. $F(x)$ cannot be set identically to zero as discussed in reference (n).

Formula 2CS—Differential in photon energy and angle.
Approximation (J), (N). Reference formula (7.2) in reference (n).

$$d\sigma_{k,\theta_0}=\frac{2Z^2r_0^2}{137}\frac{dk}{k}\frac{d\xi}{E_0^2}\{(E_0^2+E^2)(3+2\Gamma)-2E_0E(1+4u^2\xi^2\Gamma)\},$$

where

$$\xi=\frac{1}{1+u^2};\quad u=p_0\theta_0;\quad \Gamma=\ln\left(\frac{1}{\delta}\right)-2-f(Z)+\mathfrak{F}\left(\frac{\delta}{\xi}\right)$$

$$\mathfrak{F}\left(\frac{\delta}{\xi}\right)=\int_{\delta/\xi}^{\infty}\{[1-F(q)]^2-1\}\frac{[q^2-(\delta^2/\xi^2)]}{q^3}dq;\quad \delta=\frac{k}{2E_0E}.$$

Formula 2CS(a)—Complete screening $\left(\frac{\gamma}{\xi}=0,\text{ where }\gamma=\frac{100k}{E_0EZ^{\frac{1}{3}}}\right)$.

Formula 2CS with $\mathfrak{F}=\ln\left(\frac{111\gamma}{200\xi}\right)$ or $\Gamma=\ln\left(\frac{111Z^{-\frac{1}{3}}}{\xi}\right)-2-f(Z)$,

where

$f(Z)=1.2021(Z/137)^2$ for low Z
$=0.925(Z/137)^2$ for high Z. See reference (k) for further discussion.

Formula 2CS(b)—Arbitrary screening.
Formula 2CS with the form factor, $F(q)$, as an arbitrary function.

Formula 2CS(c)—Intermediate screening.
Formula 2CS with $\mathfrak{F}(\delta/\xi)$ given by

$\frac{6Z^{\frac{1}{3}}\xi}{121\delta}$	0.5	1.0	2.0	4.0	8.0	15.0	20.0	25.0	30.0	35.0	40.0	45.0	50.0	60.0	70.0	80.0	90.0	100.0	120.0
$-\mathfrak{F}(\delta/\xi)$	0.0145	0.0490	0.1400	0.3312	0.6758	1.126	1.367	1.564	1.731	1.875	2.001	2.114	2.216	2.393	2.545	2.676	2.793	2.897	3.078

Formula 2CN—Nonscreened case.
Formula 2CS with $\Gamma=\ln(1/\delta)-2-f(Z)$

Formula 3CS—Differential in photon energy.
Approximation (J). Reference Formula (1) in reference (l).

$$d\sigma_k=\frac{4Z^2r_0^2}{137}\frac{dk}{k}\left\{\left(1+\left(\frac{E}{E_0}\right)^2\right)\left[\frac{\phi_1(\gamma)}{4}-\tfrac{1}{3}\ln Z-f(Z)\right]-\left(\frac{2}{3}\frac{E}{E_0}\right)\left[\frac{\phi_2(\gamma)}{4}-\tfrac{1}{3}\ln Z-f(Z)\right]\right\}.$$

Formula 3CS(a)—Complete screening ($\gamma=0$ or $\phi_1(\gamma=0)=4\ln 183$, $\phi_2(\gamma=0)=\phi_1(\gamma=0)-\frac{2}{3}$).
Formula 3CS with $\gamma=0$.

$$d\sigma_k=\frac{4Z^2r_0^2}{137}\frac{dk}{k}\left\{\left[1+\left(\frac{E}{E_0}\right)^2-\frac{2}{3}\frac{E}{E_0}\right][\ln(183Z^{-\frac{1}{3}})-f(Z)]+\frac{1}{9}\frac{E}{E_0}\right\}.$$

Formula 3CS(b)—Arbitrary screening.

$$d\sigma_k=\frac{4Z^2r_0^2}{137}\frac{dk}{k}\left\{\left(1+\left(\frac{E}{E_0}\right)^2\right)\left[\int_\delta^1(q-\delta)^2(1-F(q))^2\frac{dq}{q^3}+1-f(Z)\right]-\tfrac{2}{3}E_0E\left[\int_\delta^1\left(q^3-6\delta^2q\ln\frac{q}{\delta}+3\delta^2q-4\delta^3\right)(1-F(q))^2\frac{dq}{q^4}+\tfrac{5}{6}-f(Z)\right]\right\},$$

where

$$\delta=k/(2E_0E).$$

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

TABLE II.—(*Continued*).

Formula 3CS(c)—Intermediate screening I ($\gamma<2$).
Formula 3CS with $\phi_1(\gamma)$ and $\phi_2(\gamma)$ given in Fig. 1.

Formula 3CS(d)—Intermediate screening II ($2<\gamma<15$).
Formula 3CS with $\phi_1(\gamma)=\phi_2(\gamma)=19.19-4\ln\gamma-4c(\gamma)$ with $c(\gamma)$ given in Fig. 2.

$$d\sigma_k=\frac{4Z^2r_0^2}{137}\frac{dk}{k}\left[1+\left(\frac{E}{E_0}\right)^2-\frac{2}{3}\frac{E}{E_0}\right]\left[\ln\frac{2E_0E}{k}-\tfrac{1}{2}-c(\gamma)-f(Z)\right].$$

Formula 3CN—Nonscreened case ($\gamma=\infty$ or $c(\gamma)=0$).
Formula 3CS with $\gamma=\infty$.

$$d\sigma_k=\frac{4Z^2r_0^2}{137}\frac{dk}{k}\left[1+\left(\frac{E}{E_0}\right)^2-\frac{2}{3}\frac{E}{E_0}\right]\left[\ln\frac{2E_0E}{k}-\tfrac{1}{2}-f(Z)\right].$$

Formula 4CS—Total radiation cross section.
Approximation (J). Reference formula (45) in reference (k).

$$\phi_{\rm rad}=\frac{4Z^2r_0^2}{137}\left[\ln(183Z^{-\frac{1}{3}})+\frac{1}{18}-f(Z)\right].$$

Formula 4CN—Total radiation cross section.
Approximation (J). Reference formula (44) in reference (k).

$$\phi_{\rm rad}=\frac{4Z^2r_0^2}{137}[\ln2E_0-\tfrac{1}{3}-f(Z)]$$

energies, the photon emission angle, and screening corrections is presented in Figs. 3–10. These graphs provide various types of theoretical intercomparisons primarily for energies above 1 Mev. Such a detailed examination of the predictions is useful only for the high-energy region where the theories are reasonably accurate and require much smaller corrections than in the low-energy region. The high-energy intercomparisons rely heavily on the extreme-relativistic predictions of Schiff[14] which depend on the validity of the complete screening approximation [see Sec. IIE (3)]. The Schiff formulas are given in a relatively simple analytical form, and have been used extensively for estimating the spectrum shape from a high-energy accelerator even though other more complicated formulas with intermediate-screening approximations are believed to be more accurate (see Table V).

(1) *Dependence of the Bremsstrahlung Spectrum on Electron Energy*

Figure 3 shows the dependence of the bremsstrahlung spectrum (integrated over the photon directions) on various initial electron kinetic energies for a platinum target ($Z=78$). The spectra for 0.05- and 0.5-Mev electrons were calculated from Formula 3BN. The spectra for 5-, 50-, and 500-Mev electrons were calculated from Formula 3BS(e). Figure 4 compares spectrum shapes predicted by Formula 3BS(e) for various electron energies.

(2) *Dependence of the Bremsstrahlung Spectrum on Photon Angle*

Figures 5(a)–(e) show the dependence of the spectrum shape on the reduced photon angle, $E_0\theta_0$, as obtained from Formula 2BS. The figures show that as the emission angle increases, the relative number of high-frequency photons increases until the trend reverses at the larger angles. For comparison, the spectrum shape integrated over the emission angle is evaluated from

[14] L. I. Schiff, Phys. Rev. **83**, 252 (1951).

TABLE III. Approximations, conditions of validity, and references for bremsstrahlung formulas of Tables I and II.

Approximation	Condition of validity
A. Nonscreened	$60Z^{-\frac{1}{3}}(1+p_0^2\theta_0^2)\gg(E_0E/k)$
B. Nonscreened	$137Z^{-\frac{1}{3}}\gg(E_0E/k)$
C. Nonscreened	$E_0\ll137Z^{-\frac{1}{3}}$
D. Complete screening	$60Z^{-\frac{1}{3}}(1+p_0^2\theta_0^2)\ll(E_0E/k)$
E. Complete screening	$137Z^{-\frac{1}{3}}\gg(E_0E/k)$
F. Complete screening	$E_0\gg137Z^{-\frac{1}{3}}$
G. Approximate screening potential:	$(Ze/r)\exp(-r/a)$
H. Born approximation	$(2\pi Z/137\beta_0)$, $(2\pi Z/137\beta)\ll1$
I. Nonrelativistic	$\beta_0\ll1$
J. Extreme relativistic	$E_0, E, k\gg1$
K. Small angles	$\sin\theta=\theta$
L. Large angles	$\theta_0\gg0$
M. Approximation in electron angle integration. Result not accurate for	$\theta_0\lesssim(Z^{\frac{1}{3}}/111E_0)$
N. Small angles	$\frac{1}{5}<p_0\theta_0<5$

[a] H. Bethe and W. Heitler, Proc. Roy. Soc. (London) **A146**, 83 (1934).
[b] H. Bethe, Proc. Cambridge Phil. Soc. **30**, 524 (1933).
[c] W. Heitler, *Quantum Theory of Radiation* (Oxford University Press, London, 1954), third edition, p. 244.
[d] F. Sauter, Ann. Physik **20**, 404 (1934).
[e] G. Racah, Nuovo cimento **11**, 469 (1934).
[f] R. L. Gluckstern and M. H. Hull, Jr., Phys. Rev. **90**, 1030 (1953).
[g] A. Sommerfeld, *Wellenmechanik* (Frederick Ungar, New York, 1950), Chap. 7.
[h] L. I. Schiff, Phys. Rev. **83**, 252 (1951).
[i] P. V. C. Hough, Phys. Rev. **74**, 80 (1948).
[j] E. Segre, *Experimental Nuclear Physics* (John Wiley & Sons, Inc., New York, 1953), p. 260.
[k] Davies, Bethe, and Maximon, Phys. Rev. **93**, 788 (1954).
[l] H. Olsen, Phys. Rev. **99**, 1335 (1955).
[m] Olsen, Maximon, and Wergeland, Phys. Rev. **106**, 27 (1957).
[n] H. Olsen and L. C. Maximon, Phys. Rev. **114**, 887 (1959).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

Formula 3BS(e) and is shown by the dashed line. In Figs. 6(a)–(e), the dependence of the cross section (Formula 2BS) on the photon emission angle, θ_0, is plotted for various photon and electron energies.

The spectrum shape integrated[15] over the photon directions with the limits from zero to a maximum value of θ_0 equal to Θ is of practical interest to experimentalists. Figures 7(a)–(e) show the spectra obtained for different values of $E_0\Theta$ by integrating Formula 2BS within the above limits of θ_0. These curves facilitate estimates of the change in thin-target spectra for different experimental arrangements that subtend various angles. In Figs. 8(a)–(c), the curves give estimates of the fraction of the total number of photons at any

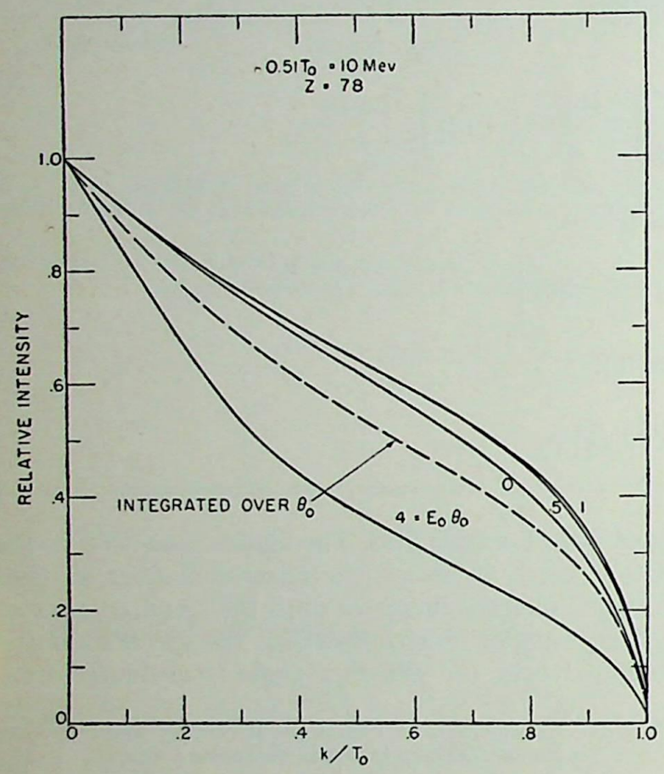


FIG. 5(a)

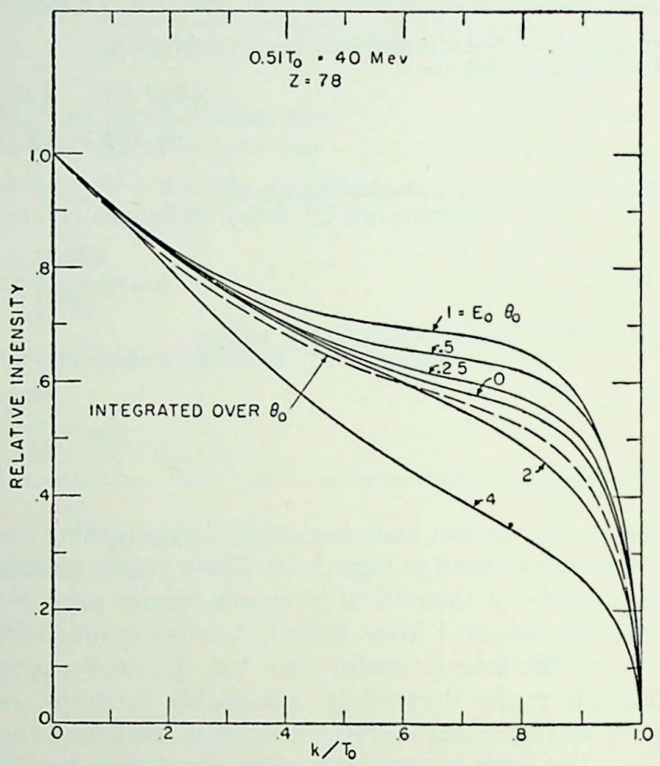


FIG. 5(c)

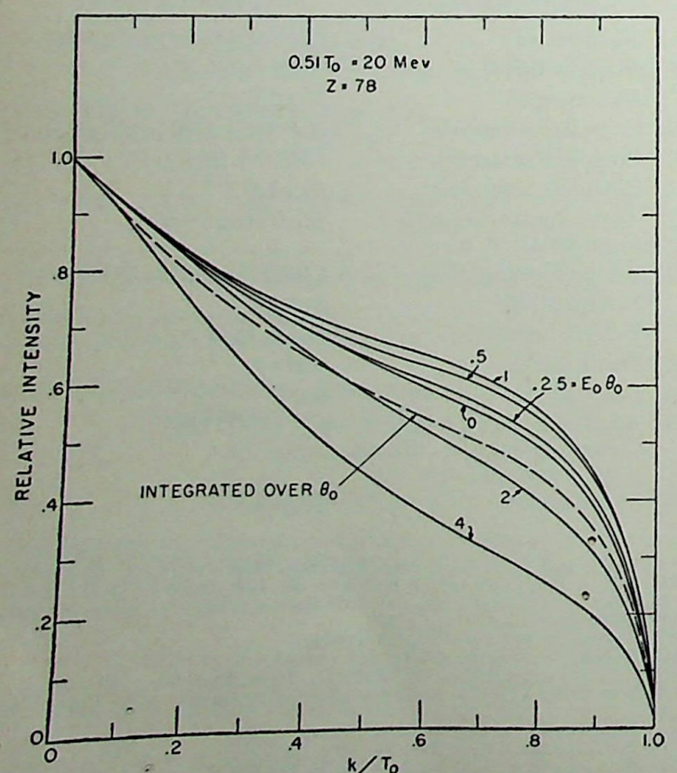


FIG. 5(b)

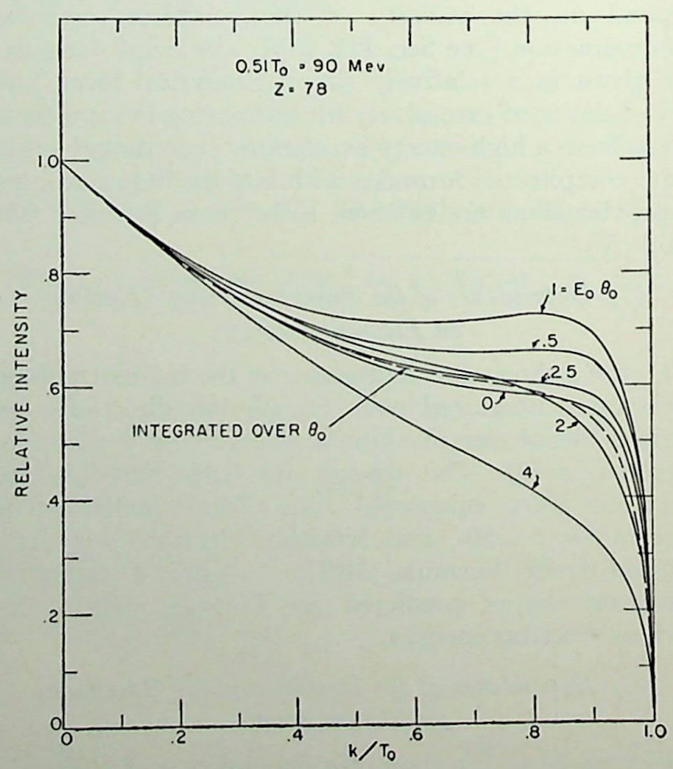


FIG. 5(d)

[15] J. H. Hubbell, J. Appl. Phys. **30**, 981 (1959).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

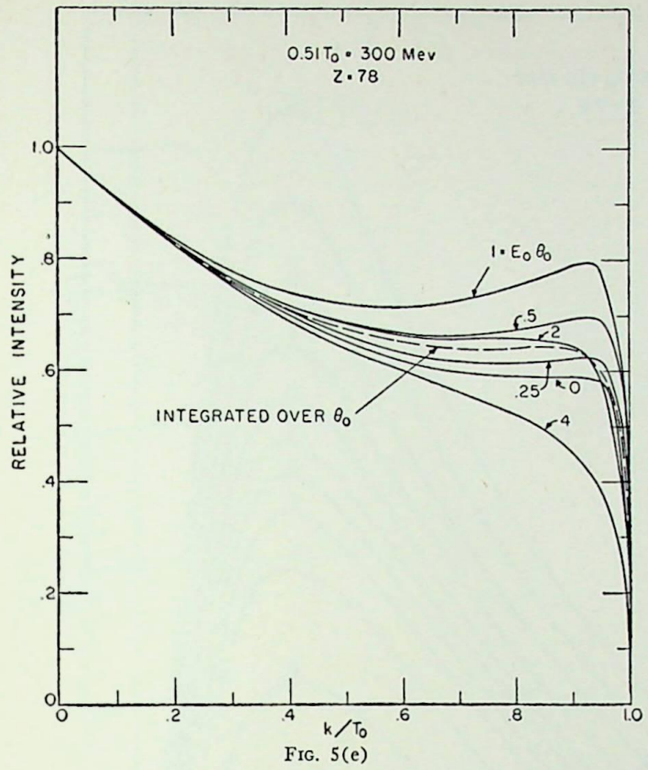


FIG. 5(e)

FIG. 5(a) Dependence of the Schiff spectrum shape on the photon emission angle, θ_0, for 10-Mev electrons and for $Z=78$. The data are obtained from Formula 2BS (solid lines) and from Formula 3BS(e) (dashed line). The values for the intensities (defined as proportional to the product of the photon energy and number per unit time) are normalized to unity at the zero photon energy. (b) Dependence of the Schiff spectrum shape on the photon emission angle, θ_0, for 20-Mev electrons and for $Z=78$. (c) Dependence of the Schiff spectrum shape on the photon emission angle, θ_0, for 40-Mev electrons and for $Z=78$. (d) Dependence of the Schiff spectrum shape on the photon emission angle, θ_0, for 90-Mev electrons and for $Z=78$. (e) Dependence of the Schiff spectrum shape on the photon emission angle, θ_0, for 300-Mev electrons.

given energy that are included within the angular limits from zero to $E_0\Theta$; these curves are obtained by graphical integration from Figs. 6(a), (c), and (e) for initial electron kinetic energies of 10, 40, and 300 Mev.

Figures 5–8 present some predictions of the Born-approximation formulas given in Table I. For comparison, the spectrum shapes as a function of the photon emission angle that are predicted by the more accurate extreme-relativistic Formula 2CS(c) in Table II, are shown in Figs. 9(a) through (e) with a normalization of unity for zero photon energy. The spectra for electron kinetic energies of 10, 20, and 40 Mev, Figs. 9(a), (b), and (c), are predicted with a zero Coulomb correction factor, $f(Z)=0$, in Formula 2CS(c), and the spectra for electron kinetic energies of 90 and 300 Mev, Figs. 9(d) and (e), include the Coulomb correction factor for $Z=78$. A comparison of the spectral shapes with and without the Coulomb correction factor shows only small differences compared to the larger effects that occur with different types of screening approximations.

(3) *Screening Effects and Coulomb Corrections*

Figures 10(a)–(e) intercompare Formula 2BS (0° Schiff), Formula 3BS(c) (Schiff's formula integrated over the photon angle), Formula 3BS (Bethe-Heitler's formula integrated over the photon angle), and the latter formula including the Davies, Bethe, Maximon correction [Formula 3CS and Sec. IIE(1)]. The three curves that are integrated over photon angle are appreciably different in shape. For example, the curves labeled "Schiff" and "Davies, Bethe, Maximon" are 10% different for an electron kinetic energy of 10 Mev at a fractional photon energy of 0.7 with the normalization used in this figure. For the sake of completeness, the spectra corrected for multiple scattering are also plotted in these figures. The multiple scattering correction as calculated by Hisdal[16] is discussed in Sec. IV.

E. Corrections for the Cross-Section Formulas

Various corrections have been obtained for the formulas given in Sec. IIC. These corrections may be classified according to three types: (1) Coulomb corrections, (2) high-frequency-limit corrections, and (3) screening corrections. In each case, the correction is restricted to a particular energy region, and is intended to apply only to the formula for a particular differential form of cross section as specified below.

(1) *Coulomb Corrections*

(*a*) *Nonrelativistic energies.*—In the nonrelativistic region where $T_0 \ll 1$, Elwert[17] has estimated a multiplicative Coulomb correction factor for the cross-section Formula 3BN(a). The Elwert factor, f_E, can be written as

$$f_E=\frac{\beta_0\{1-\exp[-(2\pi Z/137\beta_0)]\}}{\beta\{1-\exp[-(2\pi Z/137\beta)]\}}. \qquad \text{(II-6)}$$

This factor is valid only if $(Z/137)(\beta^{-1}-\beta_0^{-1})\ll 1$. This requirement forbids the use of the Elwert factor near the high-frequency limit. In addition, the Elwert correction was derived on the basis of a comparison between the nonrelativistic Born-approximation and the nonrelativistic Sommerfeld calculations. Therefore the factor is restricted to nonrelativistic electron energies. For higher electron energies (of the order of the electron rest energy), the experimental results in Sec. IIF show that the Elwert factor breaks down. As a rough guide, the Elwert factor may be expected to give results that are accurate to about 10% for electron energies below about 0.1 Mev.

(*b*) *Intermediate energies.*—In the energy region from roughly 0.1 to 2.0 Mev, Coulomb corrections to the Born-approximation formulas are not available in analytical form. Therefore these corrections must be estimated empirically from experimental results (Sec. IIF). For the cross-section formulas differential in photon energy, $d\sigma_k$, such empirical corrections cannot

[16] E. Hisdal, Phys. Rev. **105**, 1821 (1957).
[17] G. Elwert, Ann. Physik **34**, 178 (1939).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

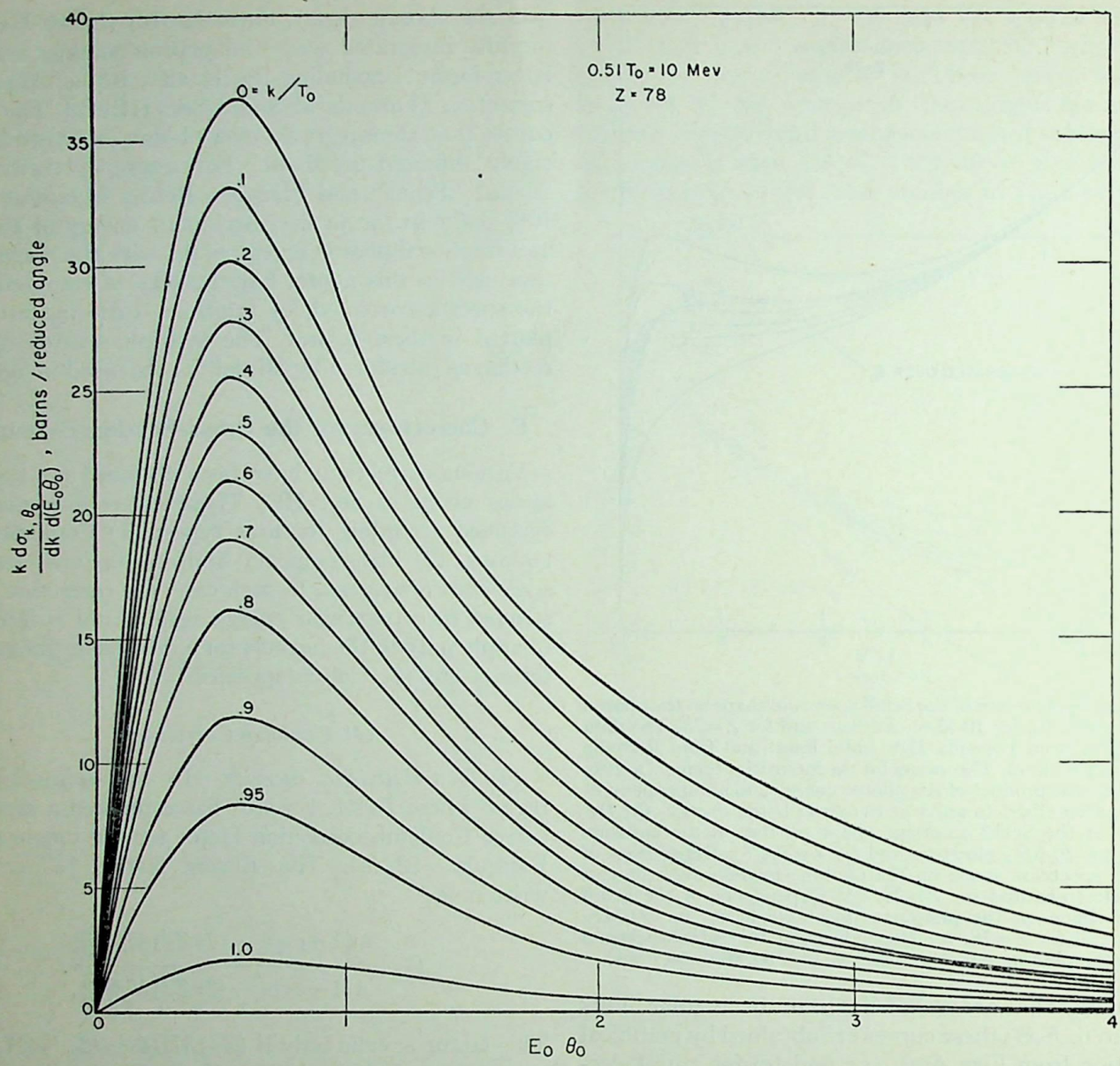


FIG. 6. Angular dependence of the Schiff cross section (Formula 2BS, $Z=78$) at various photon energies for the following electron energies: (a) 10 Mev, (b) 20 Mev, (c) 40 Mev, (d) 90 Mev, and (e) 300 Mev.

(*Figs. 6(b) to 6(e) continued on following pages*)

be determined in enough detail from the available data to cover the whole energy region. However, corrected estimates of the integrated cross section, ϕ_{rad}, are given in Sec. IIF from which empirical correction factors can be obtained. The results indicate that the corrections to the Born-approximation formulas for ϕ_{rad} are as large as a factor of two in the energy region close to the electron rest energy, and less than 10% in the energy region from about 4 to 20 Mev.

(*c*) *Extreme-relativistic energies.*—In this energy region, formulas that include the Coulomb correction for the differential cross sections $d\sigma_{k,\theta_0,\phi}$ and $d\sigma_k$ are given in Table II. A comparison of the formulas in Tables I and II shows that the Coulomb correction can be applied to the Born-approximation formulas for $d\sigma_k$ by the addition of

$$\Delta=-\frac{4Z^2r_0^2}{137}\frac{dk}{k}\left[1+\left(\frac{E}{E_0}\right)^2-\frac{2}{3}\frac{E}{E_0}\right]f(Z), \quad \text{(II-7)}$$

where $f(Z)$ is approximately equal[13] to $1.20(Z/137)^2$ for low Z and $0.925(Z/137)^2$ for high Z. This additive term is independent of the type of screening approximation that is used and is similar to the correction derived for the pair production process.[11] For lead and energies above 50 Mev, the correction decreases the Born-approximation ϕ_{rad} with intermediate screening by about 10%. The corrected cross section should be accurate to about 2%.

Accurate experimental data corroborating the cross-section values predicted by these formulas are not yet

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

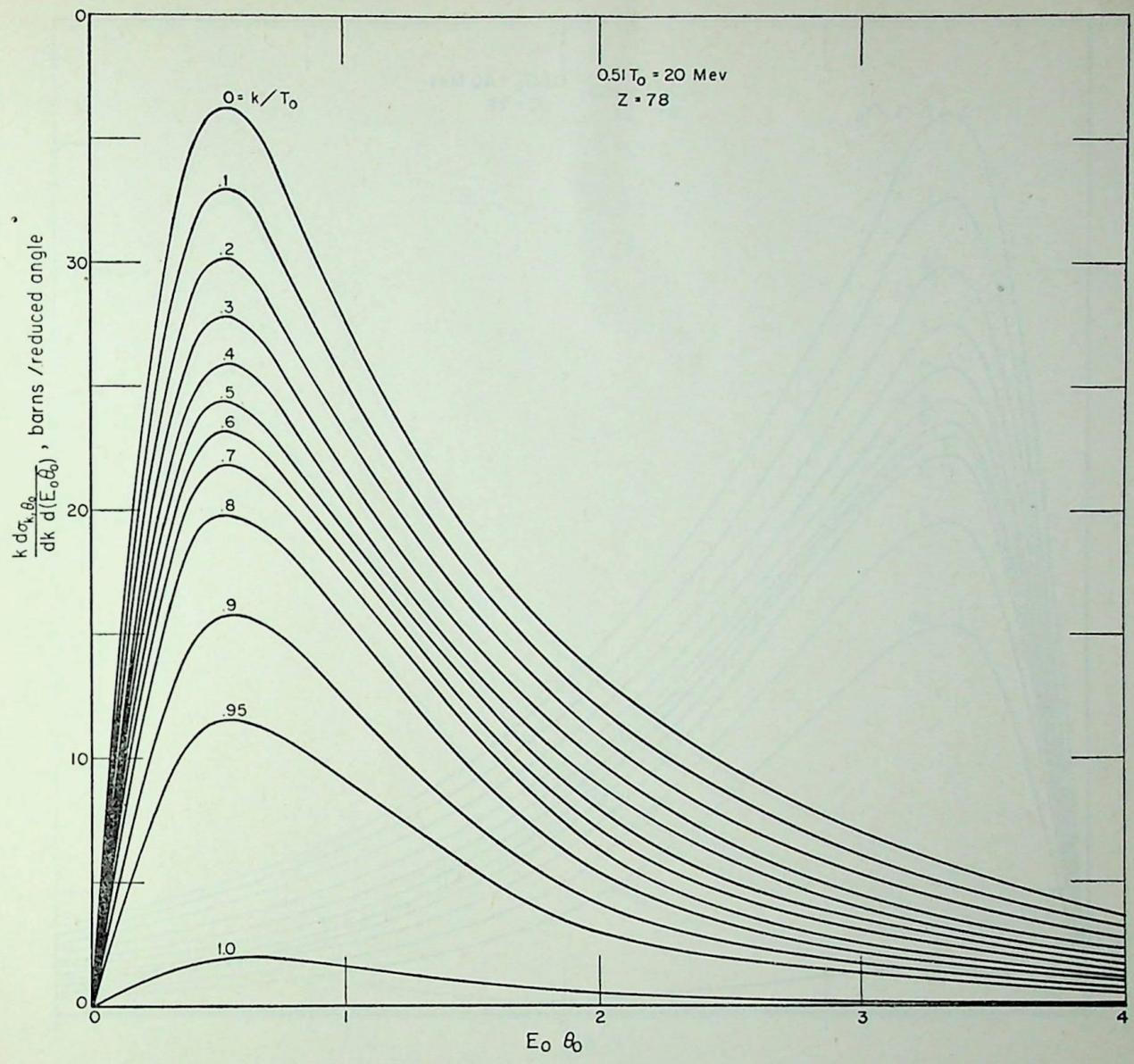


FIG. 6(b)

available. However, confirmation is available in the results of absolute pair cross sections. The ratios of the experimental pair-production cross sections to the Born-approximation cross sections are found to agree with the Davies, Bethe, Maximon[13] values, as shown in Fig. 11.

(2) *High-Frequency-Limit Corrections*

The formulas in Tables I and II are derived on the basis of certain approximations which do not permit an evaluation of the cross section at the high-frequency limit. This shortcoming has been emphasized by various experimental studies[18] which indicate that the cross section has a finite value at this limit.

Recent calculations made in the Sauter approximation (expansion in powers of $Z/137\beta_0$ and $Z/137$) by Fano[19] predict a finite value for the cross section at the high-frequency limit. In contrast, the cross section with the Born-approximation (expansion in powers of $Z/137\beta_0$ and $Z/137\beta$) becomes zero at the limit. The cross-section formulas for the high-frequency limit obtained by Fano are[20]

$$[d\sigma_{k,\theta_0}]_{k=T_0}=\frac{Z^3r_0^2}{137^2}\frac{\beta_0}{k^2E_0^3}\frac{dk}{k}\frac{4\pi\sin^3\theta_0 d\theta_0}{(1-\beta_0\cos\theta_0)^4}$$
$$\times\{1+\tfrac{1}{2}E_0(E_0-1)(E_0-2)(1-\beta_0\cos\theta_0)\} \quad \text{(II-8)}$$

and, after integration over θ_0,

$$[d\sigma_k]_{k=T_0}=4\pi\frac{Z^3}{137^2}r_0^2\frac{dk}{k}\frac{E_0\beta_0}{(E_0-1)^2}$$
$$\times\left\{\frac{4}{3}+\frac{E_0(E_0-2)}{(E_0+1)}\left[1-\frac{1}{2\beta_0E_0^2}\ln\left(\frac{1+\beta_0}{1-\beta_0}\right)\right]\right\}. \quad \text{(II-9)}$$

[18] W. C. Miller and B. Waldman, Phys. Rev. **75**, 425 (1949); Fuller, Hayward, and Koch, *ibid.* **109**, 630 (1958); D. Jamnik (private communication).

[19] U. Fano, Phys. Rev. (to be published).

[20] The cross section differential in θ_0 is proportional to $\sin^3\theta_0$, therefore, both the Sauter-approximation and the Born-approximation calculations predict that the cross section at the high-frequency limit is zero for θ_0 equal to zero.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

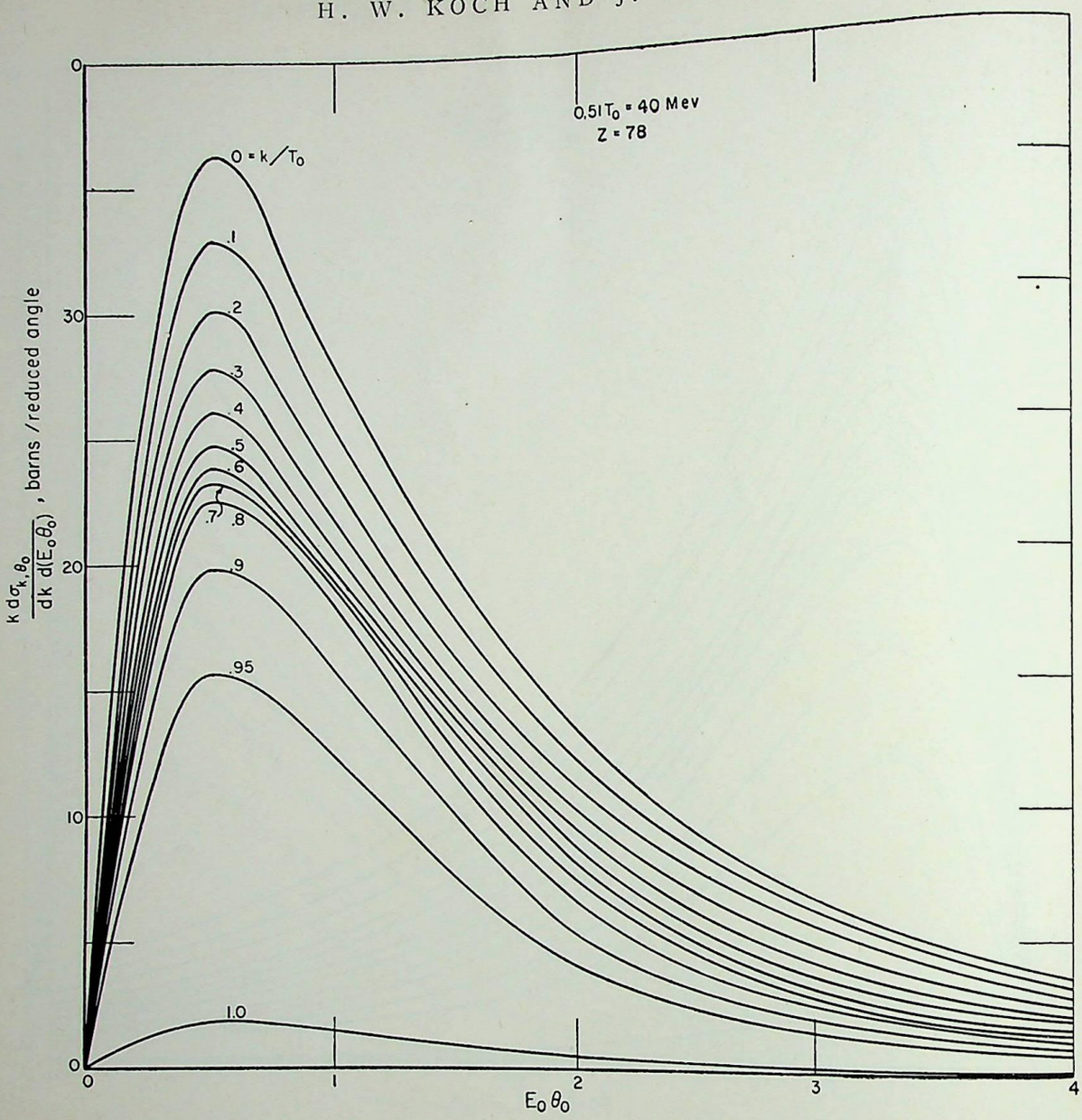


FIG. 6(c)

Fano, Koch, and Motz[21] have shown that Formula (II-9) overestimates the cross section at the high-frequency limit and that a more accurate estimate is obtained for a given electron kinetic energy if this formula is multiplied by the ratio of the "exact" to the Sauter photoelectric cross sections.[22] A summary of their results is given in Fig. 12 which shows the dependence of the bremsstrahlung cross section (integrated over photon direction) at the high-frequency limit on the incident electron energy for aluminum and gold targets. The solid lines (Sauter-Fano) are predicted by Formula (II-9) and the dashed lines (corrected Sauter-Fano) are estimated to be the corrected cross-section values. A comparison of the theoretical and experimental values indicates that the true cross sections at the high-frequency limit are predicted by the dashed curves with an accuracy of approximately 20%.

(3) *Screening Corrections*

Screening effects are most important for extreme-relativistic and for nonrelativistic electron kinetic energies. Corrections for screening have been adequately included only in the extreme relativistic calculations, where $E_0, E, k \gg 1$. The screening corrections for both the Born-approximation and Coulomb-corrected formulas in Tables I and II have been given by the Born-approximation procedure. In these screening calculations, the cross section is proportional to the square of the matrix element

$$M = \int V(r) \exp(i\mathbf{q}\cdot\mathbf{r}) d\tau, \qquad \text{(II-10)}$$

[21] Fano, Koch, and Motz, Phys. Rev. 112, 1679 (1958).

[22] Detailed formulas for the Sauter photoelectric cross section and for the "exact" cross sections of Sauter-Stobbe and Nagasaka are given by Heitler[9] and by G. White Grodstein, Natl. Bur. Standards Circ. No. 583 (1957). The "exact" photoelectric cross-section formula for high energies has been calculated recently by R. H. Pratt, thesis, University of Chicago (June, 1959), and Phys. Rev. (to be published).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

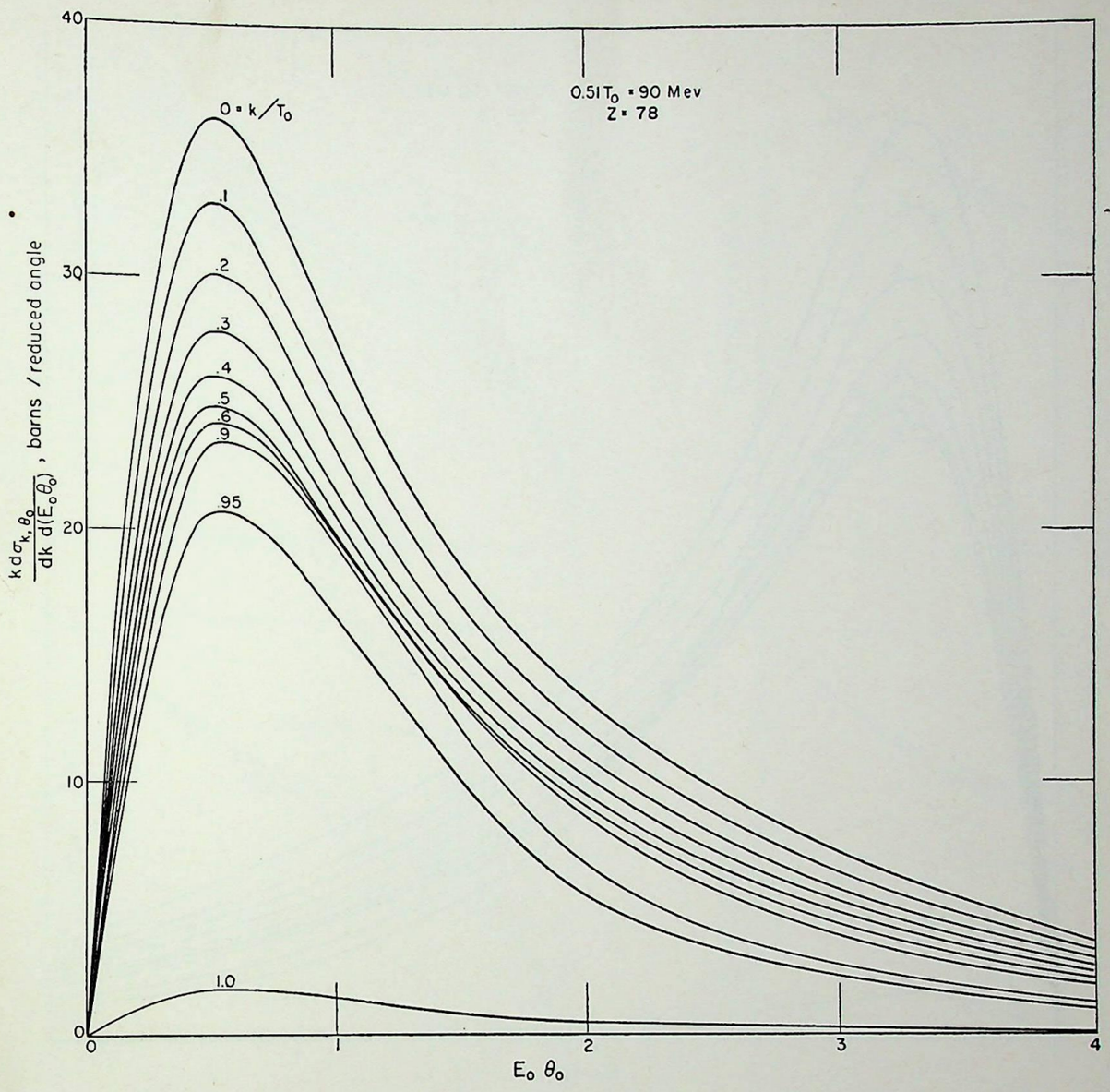


FIG. 6(d)

where $V(r)$ is the potential that determines the interaction for the bremsstrahlung process and **r** is the radius vector in units of the Compton wavelength, $\lambda\!\!\!^{-}_0$. This potential for an atom is represented by the sum, V_n+V_e, where V_n is the potential arising from the nuclear charge and V_e is the potential arising from the charge of the atomic electrons. If the atomic electron form factor is defined as

$$F_e(q,Z)=\frac{4\pi}{Ze}\int\rho(r)\left(\frac{\sin qr}{qr}\right)r^2dr, \qquad \text{(II-11)}$$

where $\rho(r)$ is the electron charge distribution, then the matrix element, M, can be written as proportional to the quantity $(Z/q^2)(F_n-F_e)$. F_n is the nuclear form factor which is roughly equal to unity.[23] Therefore, the unscreened differential cross-section formulas may be corrected for screening effects by including the multiplicative factor $[1-F_e]^2$.

[23] S. J. Biel and E. H. S. Burhop, Proc. Phys. Soc. (London) A68, 165 (1955).

For a Thomas-Fermi model, F_e depends on the quantity $qZ^{-\frac{1}{3}}$ where q has a minimum value of (p_0-p-k). At higher energies, $q_{\min}$ becomes equal to $(k/2E_0E)$ and screening calculations are expressed in terms of $\gamma=100k(E_0EZ^{\frac{1}{3}})^{-1}$. γ is approximately equal to the radius of the Thomas-Fermi atom ($r_{TF}=137Z^{-\frac{1}{3}}$) divided by $r_{\max}$, where $r_{\max}$ is the maximum impact parameter discussed by Heitler[9] and is equal to $q_{\min}^{-1}$. If $r_{\max}$ is large compared to the nuclear radius but small compared to the atomic radius r_{TF}, then γ is large and $F_e(q,Z)=0$. If $r_{\max}$ is of the order of r_{TF}, then $\gamma\sim1$ and screening must be taken into account. If the impact parameter is of the order of the nuclear radius, then the distribution of the nuclear charge must be included by a nuclear form factor[23] although the influence of the distribution of the atomic electrons can be neglected.

The dependence of $r_{\max}$ on the initial electron kinetic energy at all energies can be obtained by setting $r_{\max}$ equal to $(p_0-p-k)^{-1}$. The results are shown in Fig. 13 for k equal to $0.1T_0$, $0.5T_0$, and $0.9\,T_0$. Also, the dashed

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

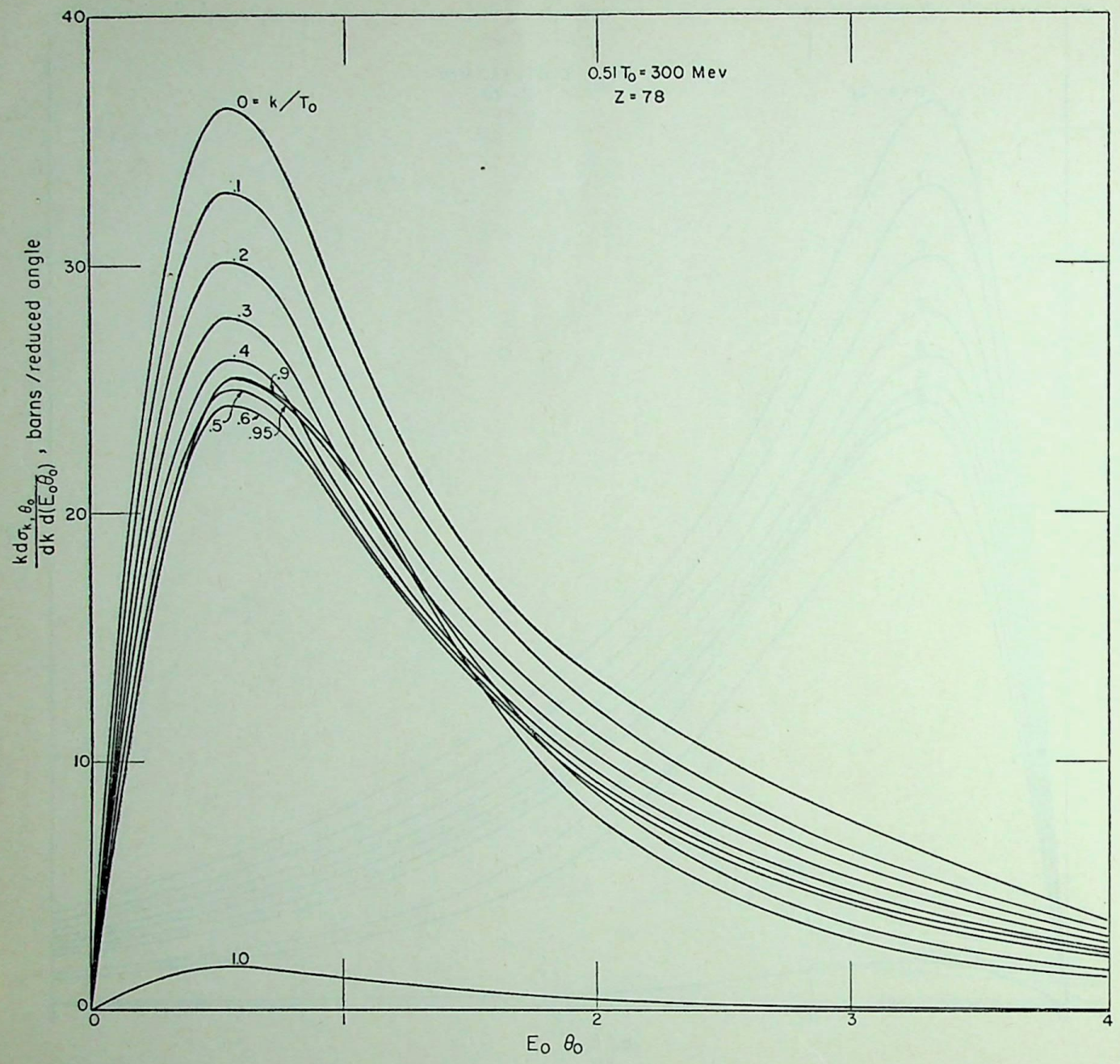


FIG. 6(e)

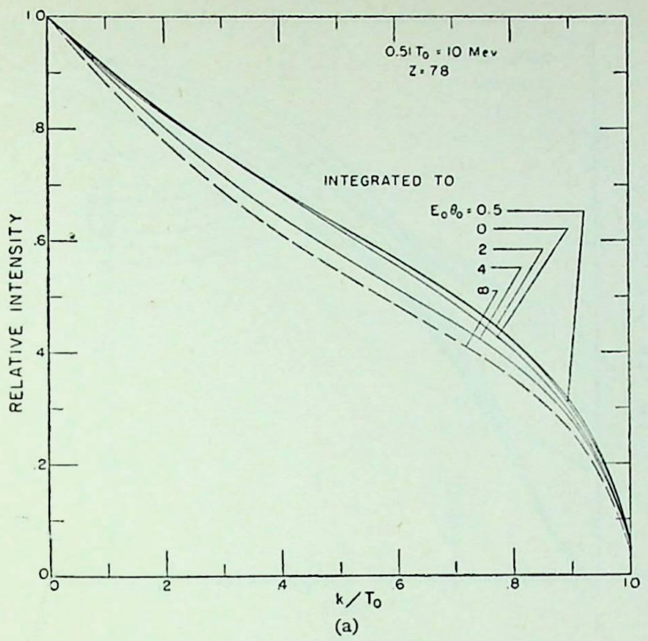
lines give the Thomas-Fermi atomic radii for beryllium and gold. A comparison of r_{max} with r_{TF} shows that r_{max} is larger than r_{TF} at low and high energies. To be specific, screening effects can be expected to become important over a large part of the spectrum for electron kinetic energies above approximately 5 Mev and below approximately 10 kev. For low fractional photon energies where $k \leq 0.1T_0$, screening effects are important for all values of T_0. It is interesting to observe that for the high photon energies the screening effects are the least important for values of T_0 approximately equal to the electron rest energy.

The accuracy obtainable with a bremsstrahlung formula corrected for screening depends on the validity of the extreme-relativistic approximations and on the adequacy of the atomic model used to provide the form factor. Only the latter will be commented on here. The most extensive calculations and applications have been based on the Thomas-Fermi atomic model. However, the Hartree self-consistent field model is more accurate[24] but more difficult to apply. The atomic form factors predicted by the two models have been compared by Nelms and Oppenheim[24] and are given in Fig. 14. The curves in this figure show that the accuracy of the Thomas-Fermi model decreases as the atomic number decreases.

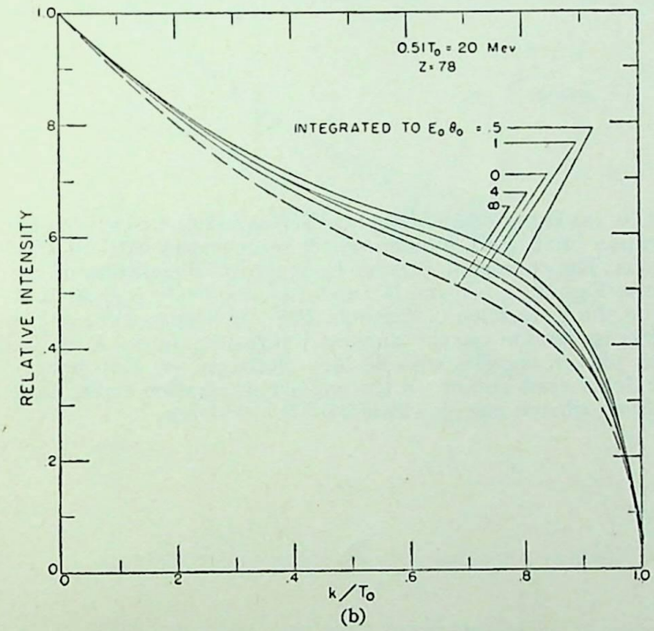
Information concerning the influence of the form factor differences on the bremsstrahlung cross section can be obtained by referring to pair production calculations. The nuclear momentum distribution in the pair production process at extreme-relativistic energies as calculated by Jost, Luttinger, and Slotnik,[25] is given in Fig. 15. Their results show that the most probable q values are of the order of 0.1. Table IV gives the ratio of the Thomas-Fermi to the Hartree form factors for representative q values, as obtained by Nelms and

[24] A. T. Nelms and I. Oppenheim, J. Research Natl. Bur. Standards **55**, 53 (1955).
[25] Jost, Luttinger, and Slotnik, Phys. Rev. **80**, 189 (1950).

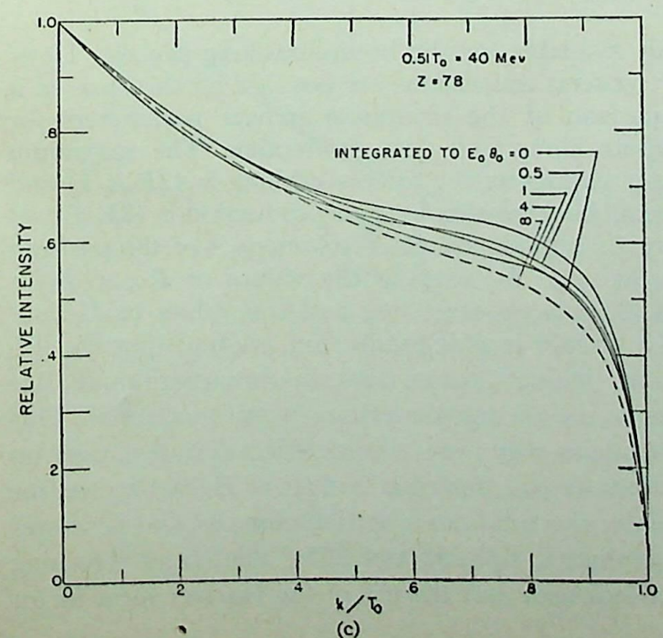
CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA



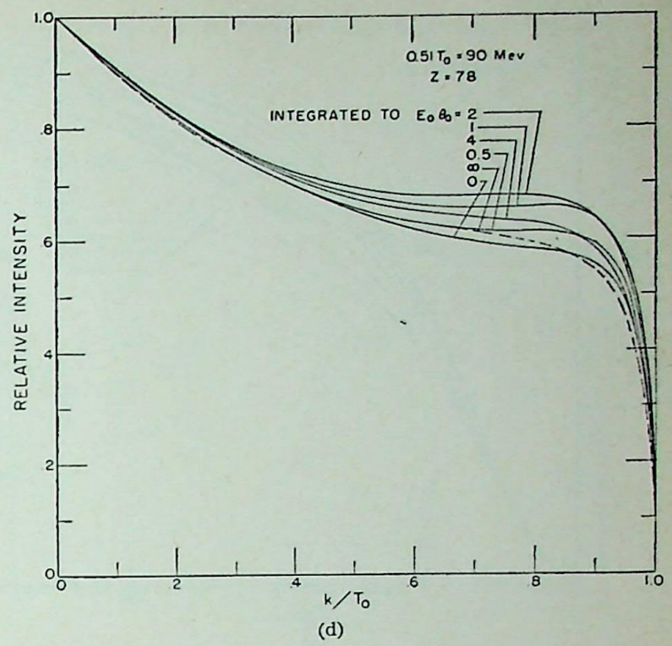
(a)



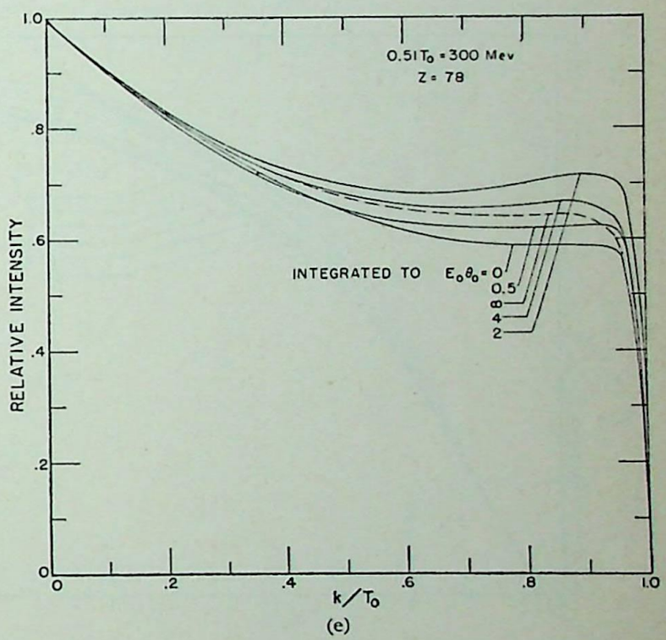
(b)



(c)



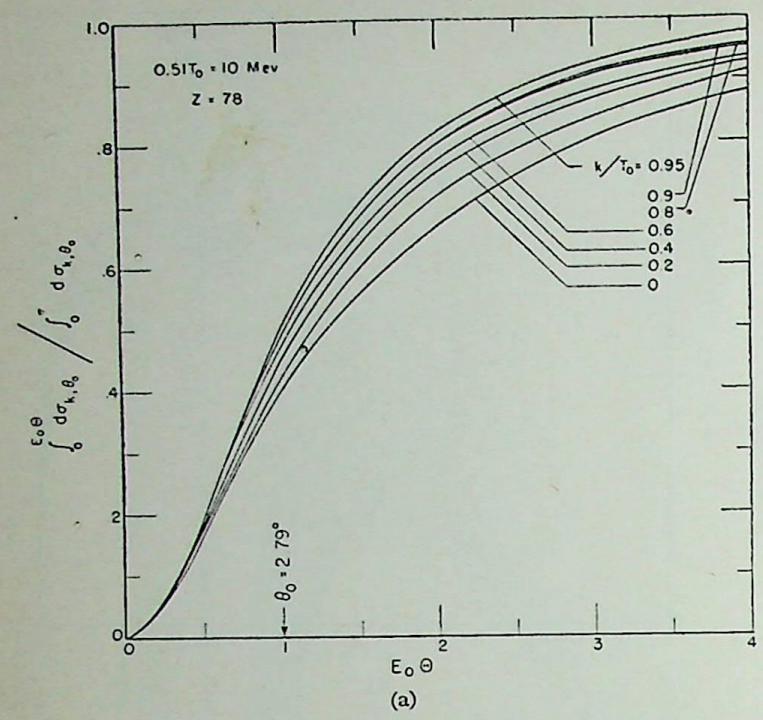
(d)

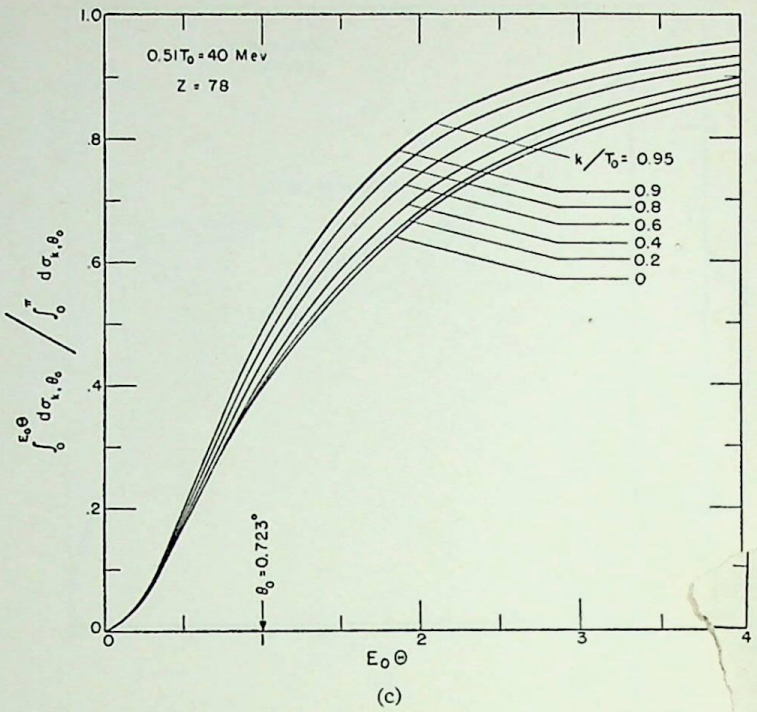


(e)

FIG. 7. (a) Comparison of Schiff spectra integrated up to various photon emission angles for 10-Mev electrons and for $Z=78$. The curves were derived by graphical integration of the cross-section curves given in Fig. 6, and may also be derived analytically by the integration of Formula 2BS. The curve labeled ∞ represents the spectrum integrated over all angles. The values of the intensity (defined as proportional to the product of the photon energy and number per unit time) are normalized to unity at the zero photon energy. (b) Comparison of Schiff spectra integrated up to various photon emission angles for 20-Mev electrons. (c) Comparison of Schiff spectra integrated up to various photon emission angles for 40-Mev electrons. (d) Comparison of Schiff spectra integrated up to various photon emission angles for 90-Mev electrons. (e) Comparison of Schiff spectra integrated up to various photon emission angles for 300-Mev electrons.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA





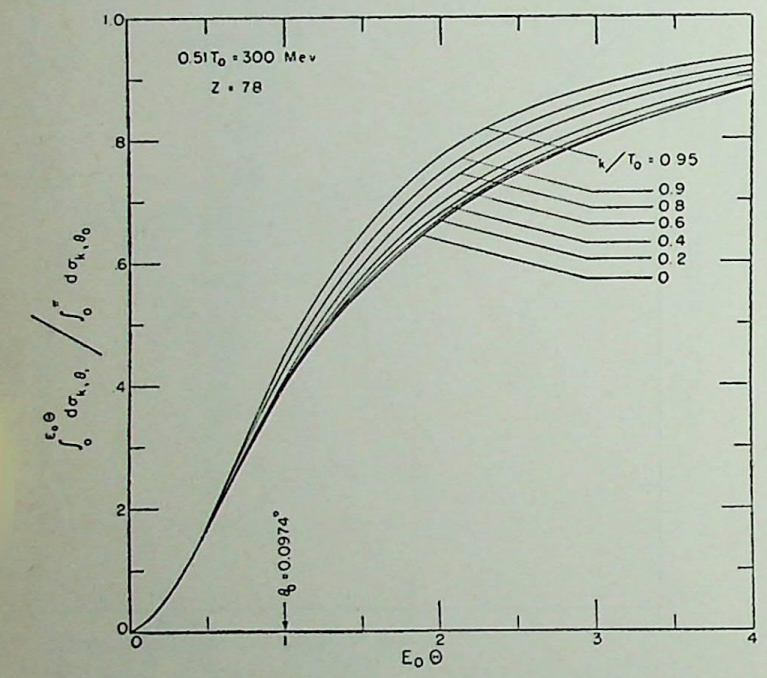


FIG. 8. (a) Dependence of the Schiff cross section on the angular integration limit, $E_0\Theta$, for various photon energies with 10-Mev electrons. The curves are derived by graphical integration of the curves in Figs. 6(a), (c), and (e), and may also be derived analytically by the integration of Formula 2BS. (b) Dependence of the Schiff cross section on the angular integration limit, $E_0\Theta$, for various photon energies with 40-Mev electrons. (c) Dependence of the Schiff cross section on the angular integration limit, $E_0\Theta$, for various photon energies with 300-Mev electrons.

Oppenheim[24] for carbon and mercury. These differences produce different estimates for the pair production cross section: for example at 60 Mev, the cross section for carbon with the Hartree form factor is increased by $\frac{1}{2}\%$ above the cross section calculated with the Thomas-Fermi form factor.[26] Comparable detailed information is not available for the bremsstrahlung process. However, general conclusions are possible on the basis of a comparison of the maximum impact parameters for bremsstrahlung and pair production. The maximum impact parameter for bremsstrahlung is $(2E_0E/k)$ and the similar expression for pair production is $(2E_+E_-/k)$ where E_+ and E_- are the total energies of the positron and electron. By varying the values of E and k for fixed E_0 in bremsstrahlung and the values of E_+ and E_- for fixed k in pair production, we find that the important impact parameters in bremsstrahlung are larger on the average than those in pair production. This fact explains why the screening effect is much larger on ϕ_{rad} than on ϕ_{pair} for equal values of E_0 and k (see, for example, the total cross sections for the two processes in reference 9, pp. 252 and 262). The larger screening effect indicates that the use of the Hartree form factor

TABLE IV. Ratios of Thomas-Fermi to Hartree atomic form factors.

q	$\left(\frac{\text{Thomas-Fermi}}{\text{Hartree}}\right)_{\text{carbon}}$	$\left(\frac{\text{Thomas-Fermi}}{\text{Hartree}}\right)_{\text{mercury}}$
0.1	1.11	0.802
0.2	2.50	1.00
0.3	3.67	1.07
0.4	4.10	1.15
0.5	4.63	1.26

[26] R. McGinnies (private communication).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

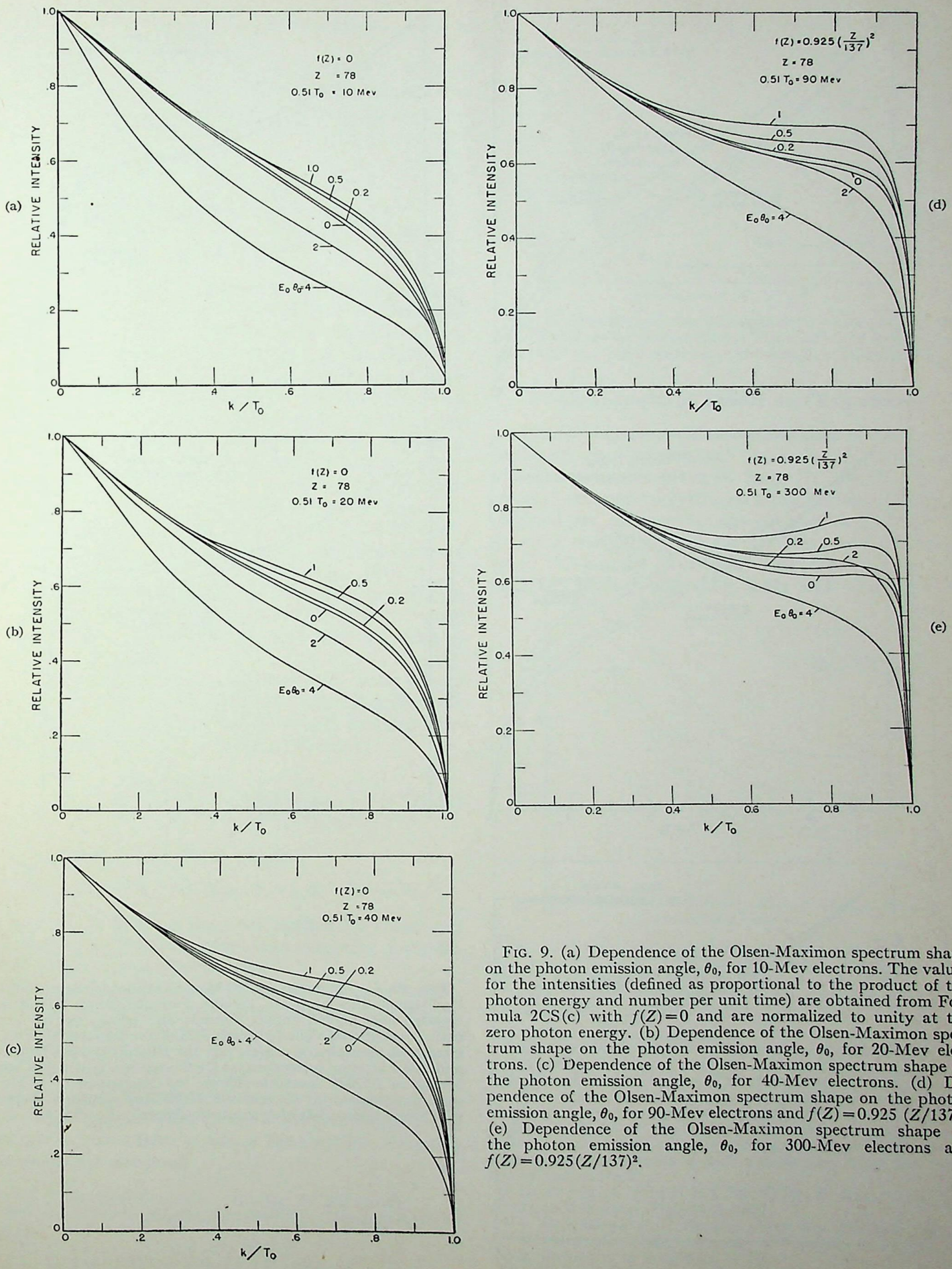


FIG. 9. (a) Dependence of the Olsen-Maximon spectrum shape on the photon emission angle, θ_0, for 10-Mev electrons. The values for the intensities (defined as proportional to the product of the photon energy and number per unit time) are obtained from Formula 2CS(c) with $f(Z)=0$ and are normalized to unity at the zero photon energy. (b) Dependence of the Olsen-Maximon spectrum shape on the photon emission angle, θ_0, for 20-Mev electrons. (c) Dependence of the Olsen-Maximon spectrum shape on the photon emission angle, θ_0, for 40-Mev electrons. (d) Dependence of the Olsen-Maximon spectrum shape on the photon emission angle, θ_0, for 90-Mev electrons and $f(Z)=0.925\ (Z/137)^2$. (e) Dependence of the Olsen-Maximon spectrum shape on the photon emission angle, θ_0, for 300-Mev electrons and $f(Z)=0.925(Z/137)^2$.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

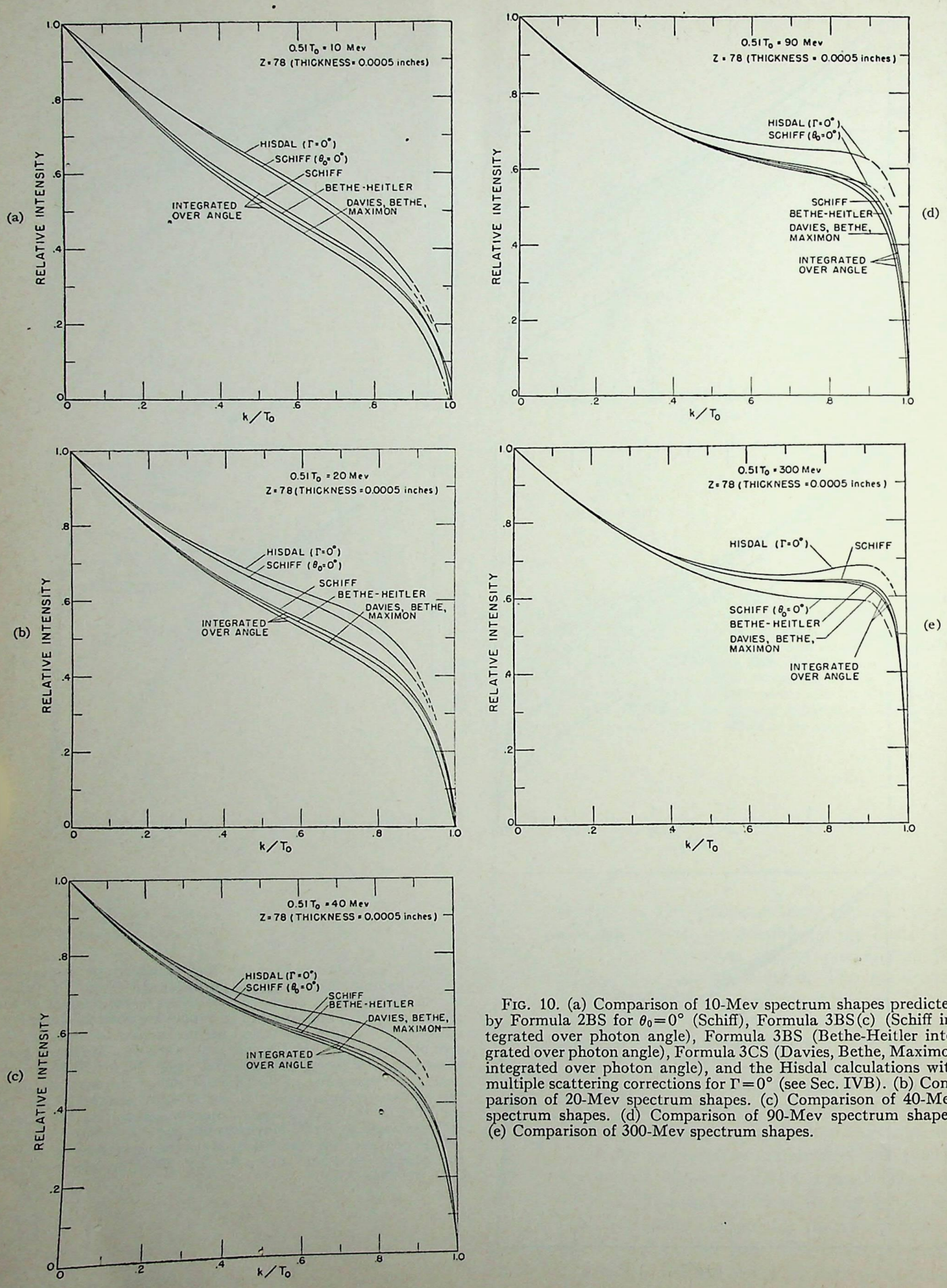


FIG. 10. (a) Comparison of 10-Mev spectrum shapes predicted by Formula 2BS for $\theta_0=0°$ (Schiff), Formula 3BS(c) (Schiff integrated over photon angle), Formula 3BS (Bethe-Heitler integrated over photon angle), Formula 3CS (Davies, Bethe, Maximon integrated over photon angle), and the Hisdal calculations with multiple scattering corrections for $\Gamma=0°$ (see Sec. IVB). (b) Comparison of 20-Mev spectrum shapes. (c) Comparison of 40-Mev spectrum shapes. (d) Comparison of 90-Mev spectrum shapes. (e) Comparison of 300-Mev spectrum shapes.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

in place of the Thomas-Fermi factor will have a greater effect on ϕ_{rad} than on ϕ_{pair}.

The first detailed study of the influence of form factors on screened bremsstrahlung cross sections was made by Bethe.[27] Bethe's calculations, which are summarized in the formulas of Tables I and II, consider four types of screening:

1, complete screening condition: $\gamma \approx 0$;
2, intermediate screening condition I: $\gamma < 2$;
3, intermediate screening condition II: $2 < \gamma < 15$;
4, no screening condition: $\gamma \gg 1$.

The Bethe[27] and Bethe-Heitler[28] screening calculations with intermediate conditions I and II were performed numerically using the tabulations of the atomic form factor for the Thomas-Fermi model given by Bethe.[29]

In the work of Schiff[14] analytical calculations were made possible by the use of the complete screening condition ($\gamma \approx 0$) and an approximate screened atom potential, V, given by $(Ze/r)\exp(-r/a)$, where $a = (111/Z^{\frac{1}{3}})$. The atomic form factor, $F_e(q,Z)$, corresponding to this potential is given by the quantity $[1+(aq)^2]^{-1}$. For many purposes the Schiff Formulas 2BS and 3BS(e) are sufficiently accurate. Schiff[14] notes that compared to the intermediate screening Formula 3BS, the complete-screening Formula 3BS(e), "is larger than it should by by less than 2% for moderate values of Z and is never more than 4% high in the worst case of large Z and energies such that screening is incomplete."

A third procedure for including form factor effects was developed by Moliére.[30] By approximating the Thomas-Fermi potential with a simple analytical expression, he obtained the following relation:

$$\frac{[1-F_e(q,Z)]}{q^2} = \sum_{i=1}^{3} \frac{\alpha_i}{\beta_i^2+q^2}, \qquad \text{(II-12)}$$

where $\alpha_1 = 0.10$, $\alpha_2 = 0.55$, $\alpha_3 = 0.35$,

$$\beta_i = Z^{\frac{1}{3}} b_i/121; \quad b_1 = 6.0, \quad b_2 = 1.20, \quad b_3 = 0.30.$$

The Moliére function has been applied by Olsen and Maximon[8] to obtain intermediate screening formulas that include Coulomb corrections.

The most accurate predictions of screening corrections to bremsstrahlung cross sections for specific target elements can be obtained by the use of the Hartree form factors in the formulas that permit the use of arbitrary form factors, e.g., Formulas 3BS(b) and 2CS(b). Unfortunately, the screening corrections for these formulas must be evaluated numerically and are not as convenient to use as the complete screening formulas just discussed.

[27] H. Bethe, Proc. Cambridge Phil. Soc. **30**, 524 (1934).
[28] H. Bethe and W. Heitler, Proc. Roy. Soc. (London) **A146**, 83 (1934).
[29] H. Bethe, Ann. Physik **5**, 385 (1930).
[30] G. Moliére, Z. Naturforsch. **2a**, 133 (1947).

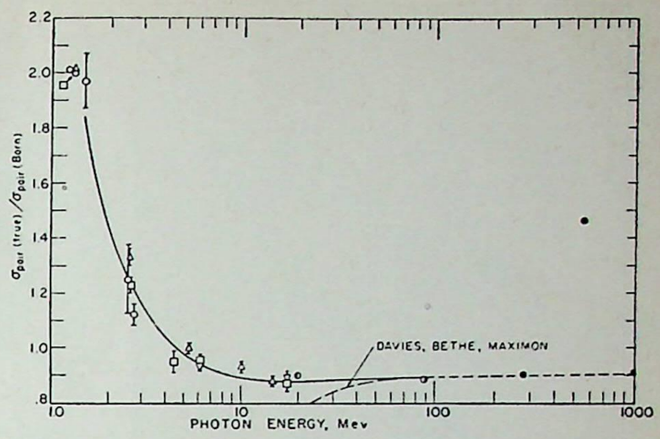


FIG. 11. Dependence of the pair cross-section ratio, $d\sigma_{pair}(\text{true})/d\sigma_{pair}(\text{Born})$, on the photon energy. The solid curve is taken from Grodstein[22] and the dashed curve is taken from reference 13.

F. Comparison of Theory and Experiment

Very few experimental determinations of the bremsstrahlung cross section are available for comparison with the estimates given in Secs. IIC, D, and E. At present, experimental data on cross sections have been obtained for electron kinetic energies of 34 kev[31] by Amrehn[32] and Röss,[32] 50 kev by Motz and Placious,[33] 90 to 180 kev by Mausbeck[34] and Zeh,[34] 0.5 and 1.0 Mev by Motz,[35] and 2.72, 4.54, and 9.66 Mev by

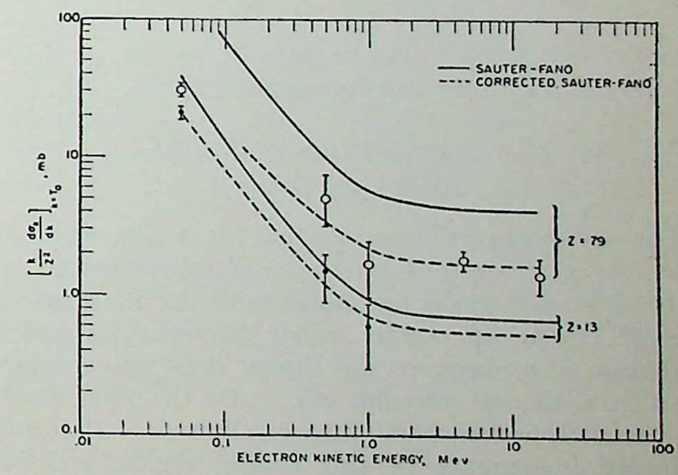


FIG. 12. Dependence of the bremsstrahlung cross section at the high-frequency limit, integrated over photon direction, on the incident electron kinetic energy. These data are obtained from reference 21, and the dashed curves are estimated to give the most accurate values for the cross section.

[31] Most of the experimental data that is available in this low-energy region has been produced by the pioneering work of Kulenkampff and co-workers. Their measurements give extensive information about relative angular distributions and spectra from thin targets, and show general agreement with the nonrelativistic Sommerfeld theory.[5] The details of their various results are not included in this report which is primarily concerned with absolute cross-section measurements and comparisons with the Born-approximation theory.
[32] H. Amrehn, Z. Physik **144**, 529 (1956); D. Röss, thesis, University of Würzburg (December, 1957).
[33] J. W. Motz and R. C. Placious, Phys. Rev. **109**, 235 (1958).
[34] H. Mausbeck, thesis, University of Würzburg (1957); H. Zeh, thesis, University of Würzburg (1957).
[35] J. W. Motz, Phys. Rev. **100**, 1560 (1955).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

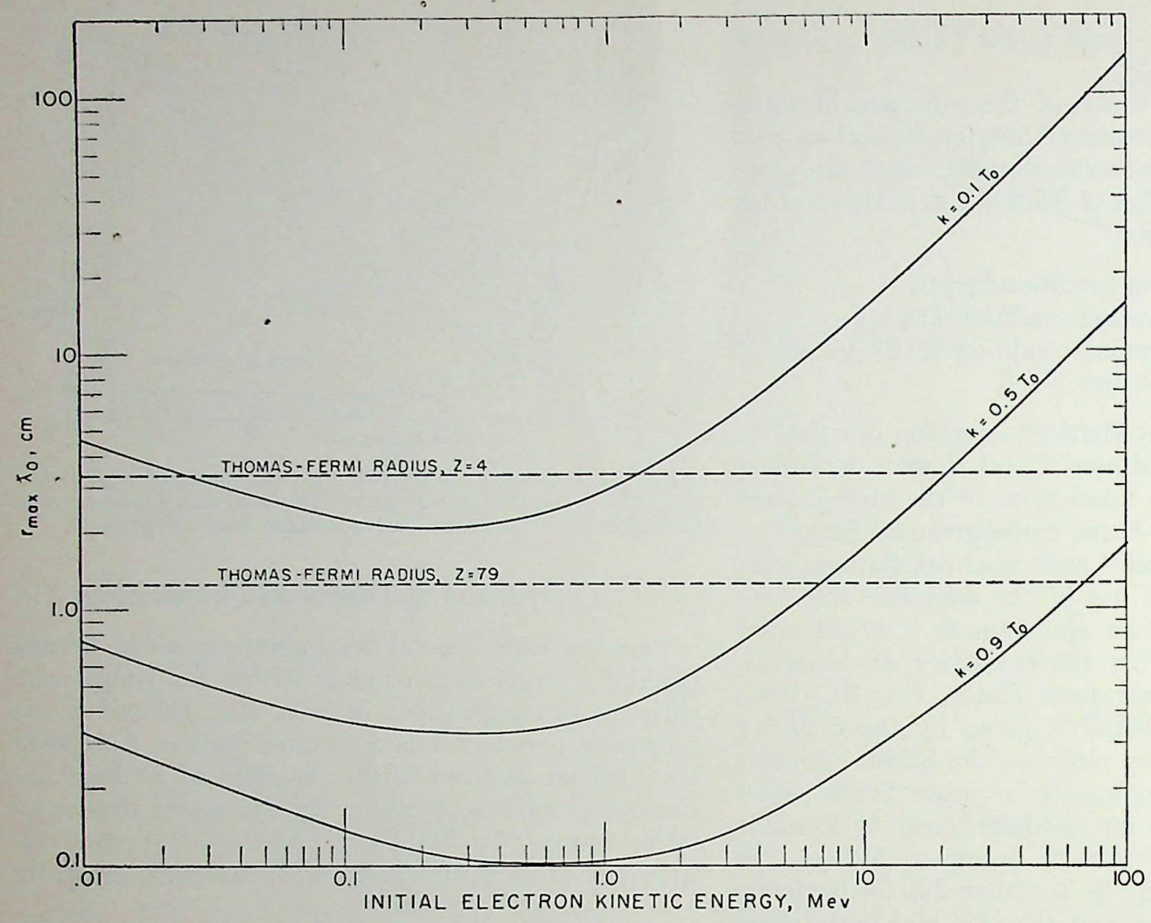


FIG. 13. Dependence of r_{max} on the initial electron kinetic energy for k equal to $0.1T_0$, $0.5T_0$, and $0.9T_0$. r_{max} is the reciprocal of the minimum momentum transferred to the nucleus ($=p_0-p-k$), and is given in units of the Compton wavelength, λ_0. The dashed lines give the values of the Thomas-Fermi radius ($=137Z^{-\frac{1}{3}}$) for beryllium and gold.

Starfelt and Koch.[36] The important results of these studies are combined and summarized below.

(1) *Cross-Section Differential in Photon Energy and Angle*

For electron kinetic energies that are small compared with the electron rest energy, the experimental results[31–33] show general agreement with the Sommerfeld theory[5] except for certain minor discrepancies which probably occur because the theory does not account for relativistic and screening effects. On the other hand, the Born-approximation theory (Formula 2BN) is seriously inadequate in this energy region and no analytical correction factors for the Born-approximation formula are available in this differential form. Furthermore, no quantitative studies are available on the importance of screening.

For electron kinetic energies that are of the same order of magnitude as the electron rest energy, the Born-approximation theory (Formula 2BN) underestimates the experimental cross section[34,35] as shown by the comparison in Fig. 16. These data also show that the differences between the theory and experiment increase with (a) the photon energy, (b) the photon angle, and (c) the atomic number of the target.

For electron kinetic energies that are large compared with the electron rest energy, the experimental results[36] agree within 10% with the Born-approximation theory. For example, in Fig. 17, the experimental cross sections for gold at 4.54 Mev[36] show general agreement with the predictions of the screened, extreme-relativistic Schiff Formula 2BS and of the unscreened Sauter Formula 2BN. There are differences in detail (generally less than 10% in this energy region): (a) near the high-frequency limit, the experimental cross sections are greater than the Schiff cross sections which in turn are greater than the Sauter cross sections; (b) in the low-frequency region, the experimental cross sections show good agreement with the Schiff cross sections, but are less than the unscreened Sauter cross sections. For low Z targets, there is better agreement with the Sauter formula.

(2) *Cross-Section Differential in Photon Energy*

A comparison[21] of experimental and theoretical values for the cross section differential in photon energy, $d\sigma_k$, is given in Figs. 18–21 for electron energies of 0.05, 0.5, 1.0, and 4.5 Mev. Each of these figures gives the estimates of (a) the Born-approximation cross sections [Formulas 3BN or 3BN(a)]; (b) the corrected Sauter-Fano cross sections at the high-frequency limit [Sec. IIE(2)]; (c) the Elwert-Born approximation cross sections [Sec. IIE(1)]; and (d) the experimental results. The solid lines show the cross sections computed from Formula 3BN(a) for 0.05 Mev, and Formula 3BN for

[36] N. Starfelt and H. W. Koch, Phys. Rev. **102**, 1598 (1956).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

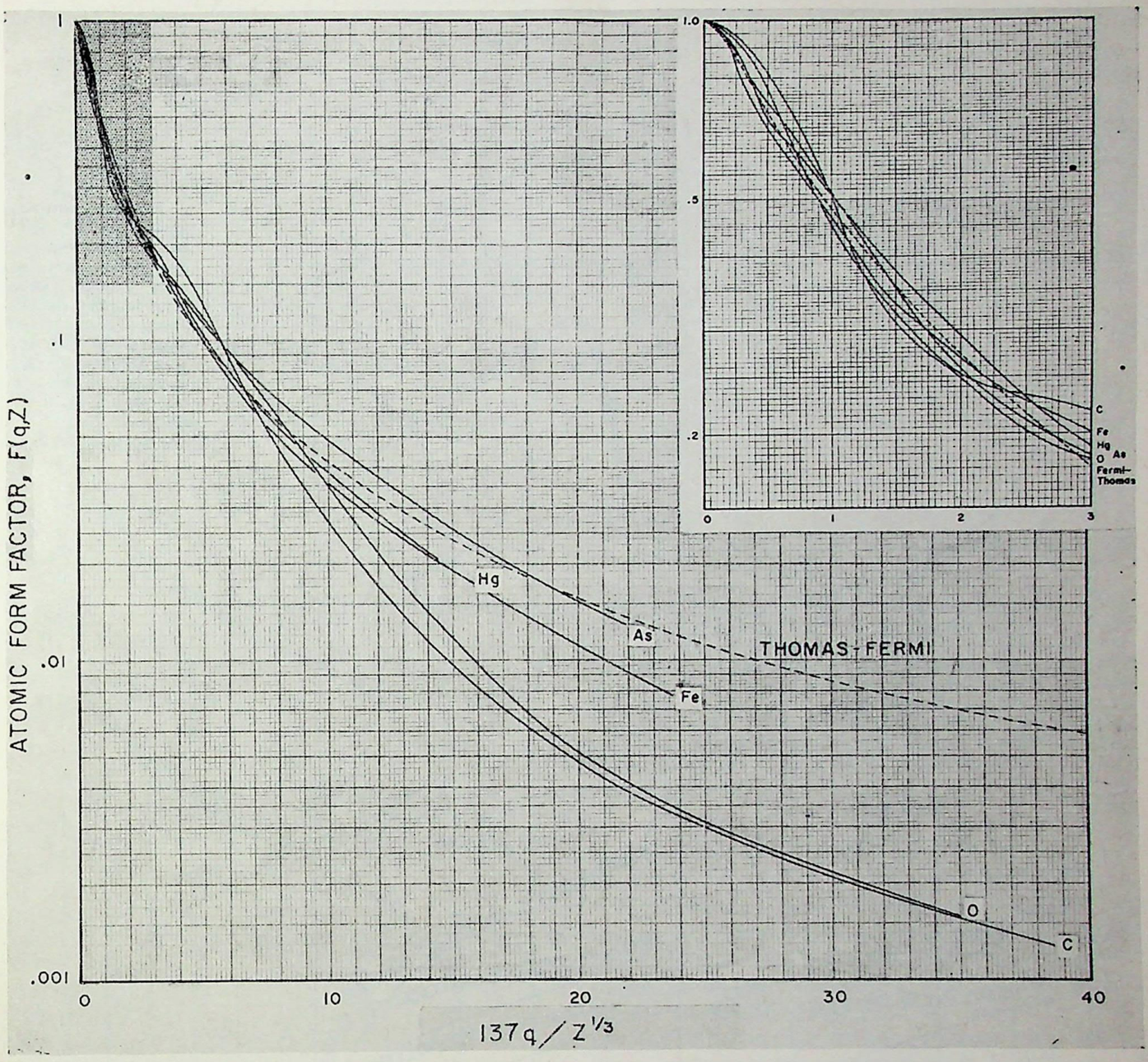


FIG. 14. Evaluation[24] of the atomic form factor, $F(q,Z)$, for the Hartree self-consistent field model (solid lines) and for the Thomas-Fermi model (dashed line), as a function of the nuclear momentum transfer, q.

0.5, 1, and 4.5 Mev. The dashed lines give the Born-approximation cross sections corrected by the Elwert factor defined in Formula (II-4). The comparison with the experimental results shows that the Elwert correction gives the most accurate results at very low energies (below 0.1 Mev). For electron kinetic energies of the order of the electron rest energy, the cross sections obtained with the Elwert correction factor are still less than the experimental values (by as much as a factor of two in the worst case). For very high energies, the Born-approximation theory overestimates the actual cross sections, and the Elwert factor is no longer valid, although it gives good agreement with experiment in the 5-Mev cross-over region [see Sec. IIF(3)]. The corrected Sauter-Fano cross sections at the high-frequency limit show good agreement with the experimental results as noted previously in Fig. 12.

(3) *Total Cross Section*

The experimental values for the total cross section, ϕ_{rad} (defined in Sec. IIC), are shown in Fig. 22 by the closed and open circles for initial electron kinetic energies of 0.05 Mev,[33] 0.5 Mev,[35] and 1.0 Mev.[35] The theoretical values are shown by the solid lines, which are predicted by Formulas 4BN(a) in the region where $T_0 \leq 0.5$ and by Formula 4BN for no screening. The curves that include screening corrections for $Z=13$ and 79 are obtained by numerical integration of the intermediate screening Formulas 3BS(c) and 3BS(d). For

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

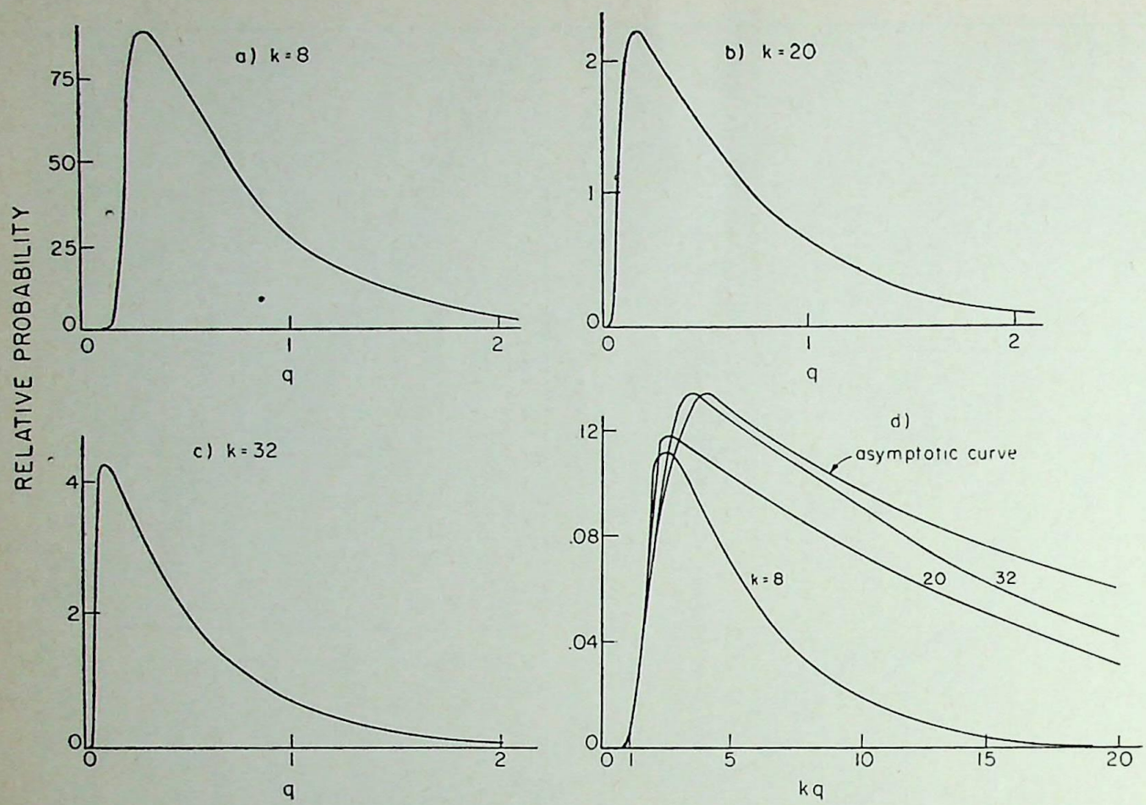


FIG. 15. Momentum distributions[25] for the recoil nucleus in nuclear pair production for several photon energies. The relative number of recoils are plotted in Figs. 15(a), (b), and (c) as a function of the momentum, q. These curves are summarized in (d) and are compared with the asymptotic curve.

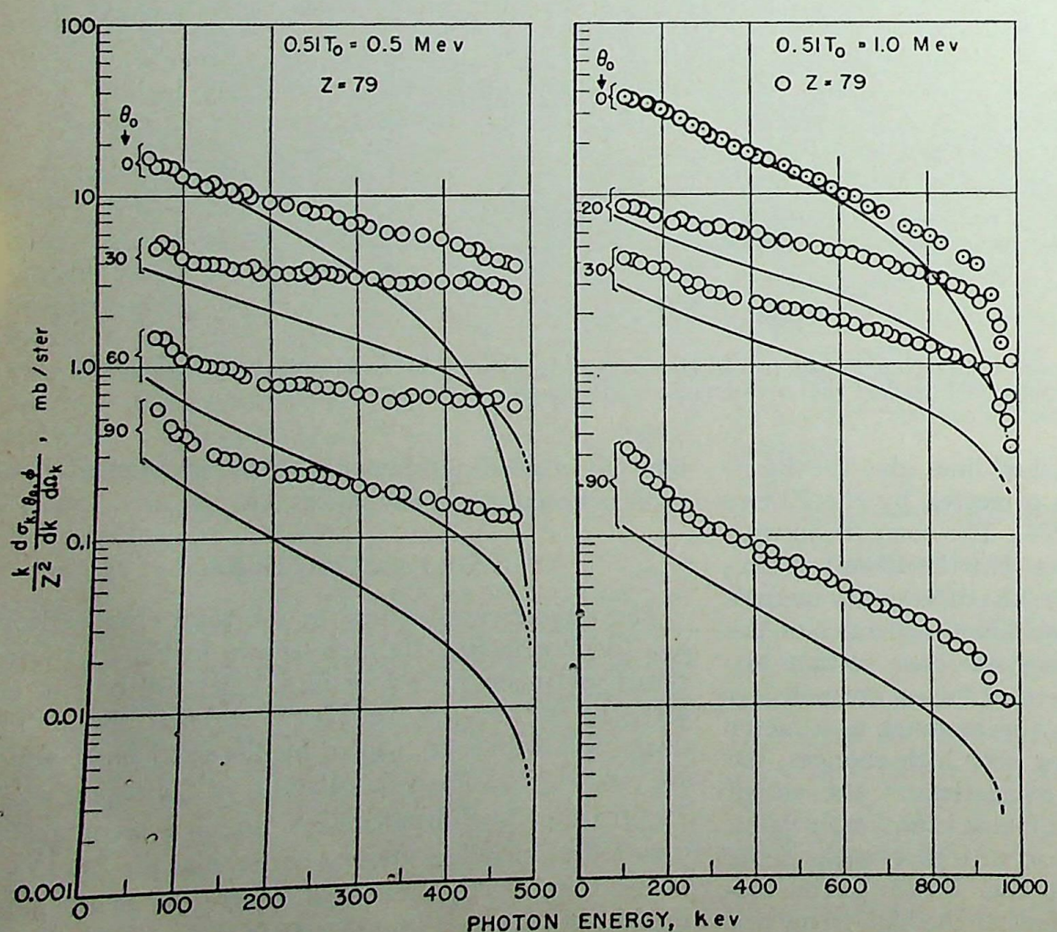


FIG. 16. Dependence of the bremsstrahlung cross section $d\sigma k,\theta_0,\phi$ on photon energy and angle, θ_0, for 0.5- and 1.0-Mev electrons. The theoretical cross sections shown by the solid curves are calculated from Formula 2BN, and the experimental values[35] for gold are given by the open circles.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

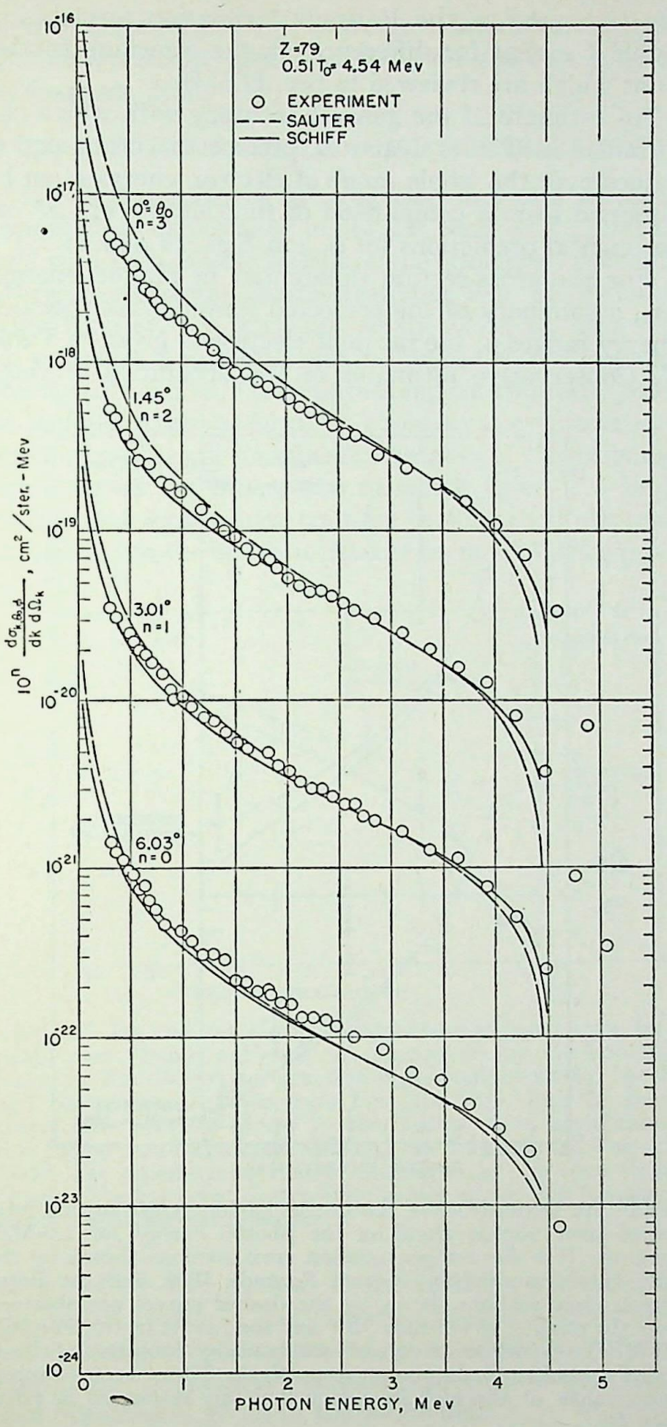


FIG. 17. Dependence of the bremsstrahlung cross section $d\sigma_{k,\theta_0,\phi}$ for gold, on the photon energy and angle, θ_0, for 4.54-Mev electrons. The theoretical cross sections are given by the solid curve (Schiff, Formula 2BS), and by the dashed curve (Sauter, Formula 2BN). The experimental values[36] for gold are given by the open circles.

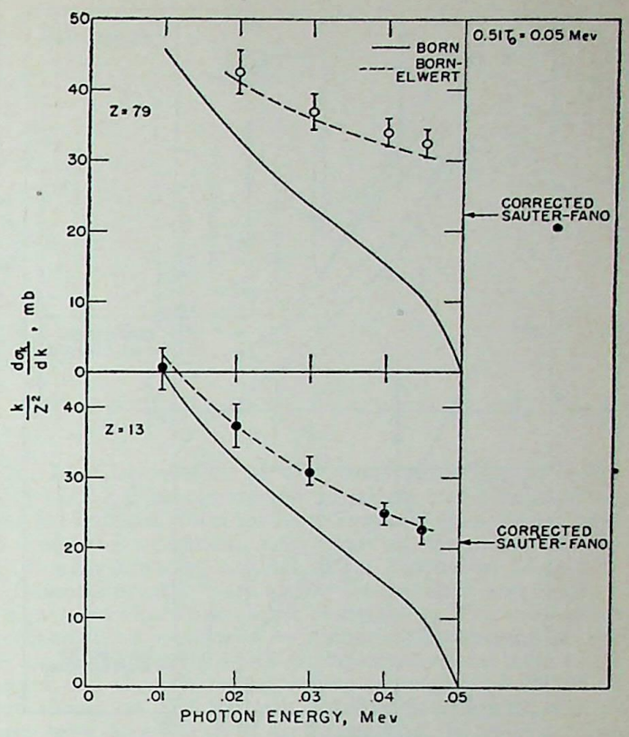


FIG. 18. Dependence of the bremsstrahlung cross section integrated over photon angle on the photon energy for 0.05-Mev electrons. The Born-approximation cross sections shown by the solid curves are calculated from Formula 3BN(a), and the Born-Elwert cross sections shown by the dashed curves are obtained from the product of Formula 3BN(a) and the Elwert factor, Formula (II-4). The experimental values[33] are shown by the open and closed circles for gold and aluminum, respectively. The corrected Sauter-Fano values at the high-frequency limit are estimated in reference 21.

extreme-relativistic energies, the triangles give the most accurate theoretical cross-section values[37] for $Z=79$, which are estimated by numerical integration of the Coulomb-corrected Formula 3CS. The most accurate values for ϕ_{rad}, which are estimated from the above combined data, are shown by the dashed curve for $Z=13$ and the dot-dashed curve for $Z=79$.

Approximate correction factors for the Born-approximation ϕ_{rad} values with screening have been estimated as a function of the initial electron kinetic energy from the ratios of the cross-section values shown in Fig. 22 by the empirical (dashed and dot-dashed) curves and by the Born-approximation curves with screening. These estimated factors are given in Fig. 23, and show that the ratios are equal to unity at the energy of approximately 10 Mev for aluminum and 6 Mev for gold. For energies larger than these crossover energies, the Born-approximation formula overestimates the cross section; for smaller energies, the reverse is true. Even though the correction factor for ϕ_{rad} may be close to unity for a given electron energy, much larger (or smaller) correction factors may be required for the differential cross section, $d\sigma_k$, in the high- or low-frequency region because of a crossover effect (see Fig. 39 in Bethe and Salpeter, reference 1), which is masked by the integration of $d\sigma_k$. It is interesting to observe in Fig. 23, that the maximum correction factor for the Born-approximation calculations occurs at electron kinetic energies approximately equal to the electron rest energy.

[37] For $Z=13$, the corrected values for energies above 50 Mev have only small differences (less than one percent) with the values shown by the solid line for $Z=13$.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

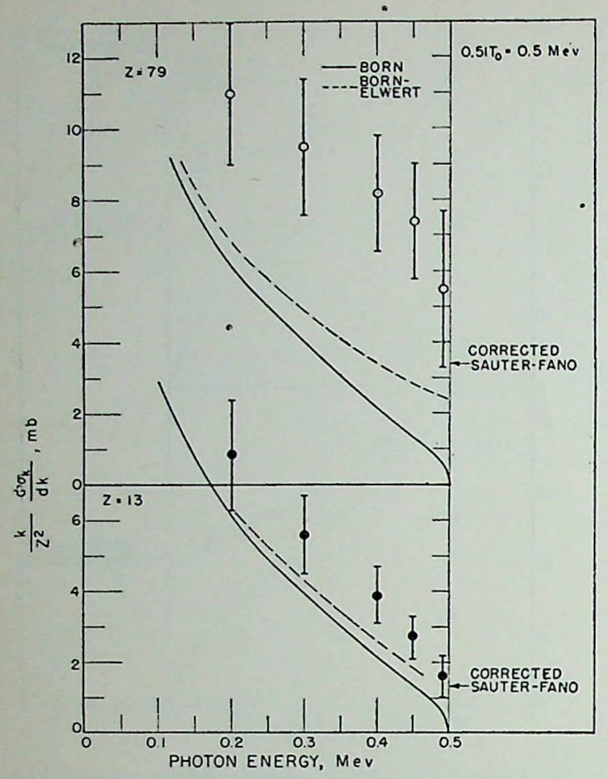


FIG. 19. Dependence of the bremsstrahlung cross section integrated over photon angle on the photon energy for 0.5-Mev electrons. The Born-approximation cross sections shown by the solid curves are calculated from Formula 3BN, and the Born-Elwert cross sections shown by the dashed curves are obtained from the product of Formula 3BN and the Elwert factor, Formula (II-4). The experimental values[35] are shown by the open and closed circles for gold and aluminum, respectively. The corrected Sauter-Fano values at the high-frequency limit are estimated in reference 21.

G. Summary

A survey of the foregoing data leads to some general conclusions about the accuracy of the cross-section values predicted by the various formulas and correction factors. Also, suggestions can be made for selecting formulas that give the best estimates for the cross section or that can be easily evaluated to give reasonably accurate results. These judgments are summarized in the following.

In Table I, the screened formulas depend on the extreme-relativistic approximation and therefore are valid only in the energy region $T_0 \gg 1$. For $T_0 \ll 1$, only the nonscreened formulas are applicable.[38] The nonscreened formulas require relatively large correction factors except in the region near the crossover energy (see Fig. 22). At the extreme relativistic energies the nonscreened formulas are less accurate than the screened formulas.

In Table II, the extreme-relativistic cross section formulas for $d\sigma_{k,\theta_0,\phi}$ and $d\sigma_k$ are estimated to have an accuracy that is given approximately by the factor $(Z/137)^2(\ln E/E)$. For $f(Z)=0$, the formulas in Table II are the same as the Born-approximation formulas in Table I except for differences in the screening corrections which are reviewed in Sec. IIE(3).

An estimate of the general accuracy with which the formulas in Tables I and II predict the cross-section values over the whole range of electron energies can be obtained from a comparison of the theoretical and experimental predictions for ϕ_{rad} in Figs. 22 and 23.

For the cross-section differential in photon energy, $d\sigma_k$, a summary of the corrected formulas for specified energy ranges of the incident electron is given in Table V. Conservative estimates of the accuracies of these formulas have been made on the basis of the experimental data assembled in this report. The greatest uncertainties are in the energy range from 0.10 to 2.0 Mev. Because of the uncertainties of screening effects, no corrected formulas are given for the energy region below 0.1 Mev. These corrected formulas are tentative and it can be expected that some will be replaced by more accurate expressions as more data becomes available.

For the cross-section formulas differential in photon energy and angle, $d\sigma_{k,\theta_0,\phi}$, no quantitative corrections are available for low and intermediate energies because

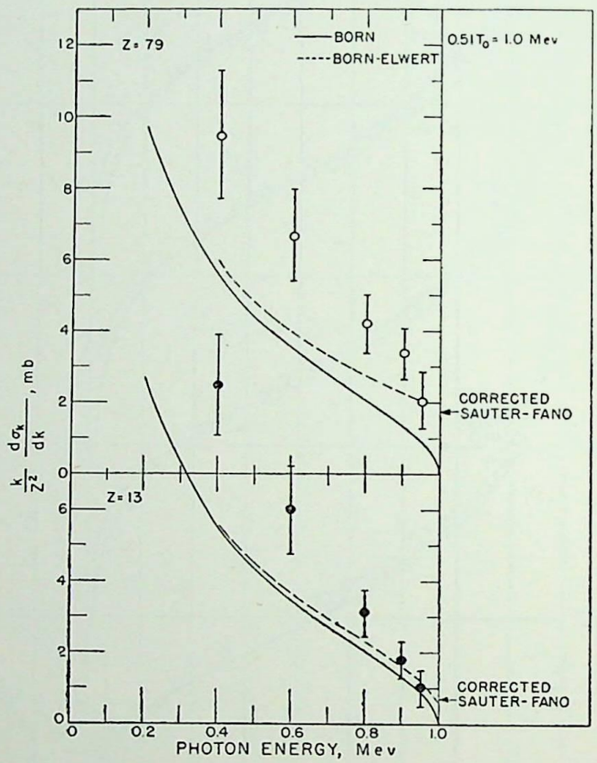


FIG. 20. Dependence of the bremsstrahlung cross section integrated over photon angle on the photon energy for 1.0-Mev electrons. The Born-approximation cross sections shown by the solid curves are calculated from Formula 3BN, and the Born-Elwert cross sections shown by the dashed curves are obtained from the product of Formula 3BN and the Elwert factor, Formula (II-4). The experimental values[35] are shown by the open and closed circles for gold and aluminum, respectively. The corrected Sauter-Fano values at the high-frequency limit are estimated in reference 21.

[38] A Born-approximation formula that includes screening effects without the extreme-relativistic approximation, has been given by Gluckstern and Hull, Phys. Rev. 90, 1030 (1953). This formula applies mainly to the low-frequency region and has been found to be inadequate[36] in the high-frequency region.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

of insufficient data. For extreme relativistic energies, the most accurate estimates (3%) for $d\sigma_{k,\theta_0\phi}$ are given by Formula 2CS.

III. ELECTRON-ELECTRON BREMSSTRAHLUNG

The bremsstrahlung cross-section formulas for electron-nuclear interactions in Sec. IIC vary as Z^2. For targets with high atomic numbers, the additional influence of electron-electron bremsstrahlung can be included approximately by replacing Z^2 by $Z(Z+1)$. However for very low Z elements such as hydrogen or beryllium, the electron-electron bremsstrahlung contributions must be included more accurately. Cross-section calculations for this process are complicated because of the exchange character of the interaction in which there is a large energy and momentum transfer to the recoil electron, in contrast to the electron-nuclear bremsstrahlung process in which the nucleus is assumed to be infinitely heavy. No complete calculations are available for predicting the detailed features of electron-electron bremsstrahlung.[39] A summary of pertinent results that have been obtained is given in the following.

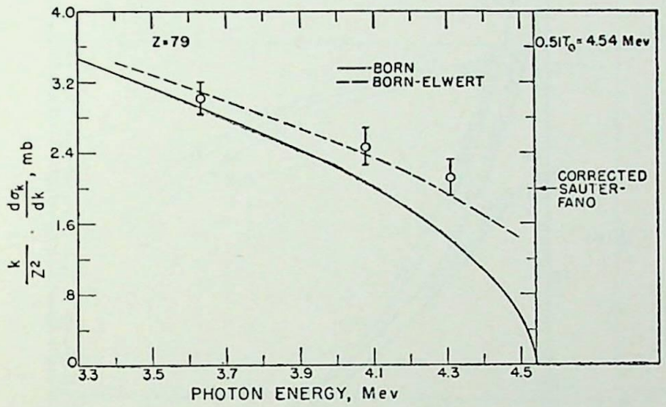


FIG. 21. Dependence of the bremsstrahlung cross section integrated over photon angle on the photon energy for 4.54-Mev electrons. The Born-approximation cross sections shown by the solid curve are calculated from Formula 3BN, and the Born-Elwert cross sections shown by the dashed curve are obtained from the product of Formula 3BN and the Elwert factor, Formula (II-4). The experimental values[36] are shown by the open circles for gold. The corrected Sauter-Fano values at the high-frequency limit are estimated in reference 21.

A. Maximum Photon Energy

In the electron-electron bremsstrahlung process, the maximum photon energy that is available in the laboratory system at the laboratory angle θ_0 is[35]

$$k_{\max}=F/(1-\sqrt{F}\cos\theta_0), \qquad \text{(III-1)}$$

where F is equal to $(E_0-1)/(E_0+1)$. Table VI gives some values of $k_{\max}$ at zero and 90 degrees obtained from Formula (III-1) for various incident electron kinetic energies. From the very sparse experimental information[35,36] available on electron-electron bremsstrahlung, some results[35] have shown reasonably good agreement with the values of $k_{\max}$ predicted by Formula (III-1).

[39] For a general review of the available theories on electron-electron bremsstrahlung, see J. Joseph and F. Rohrlich, Revs. Modern Phys. **30**, 354 (1958).

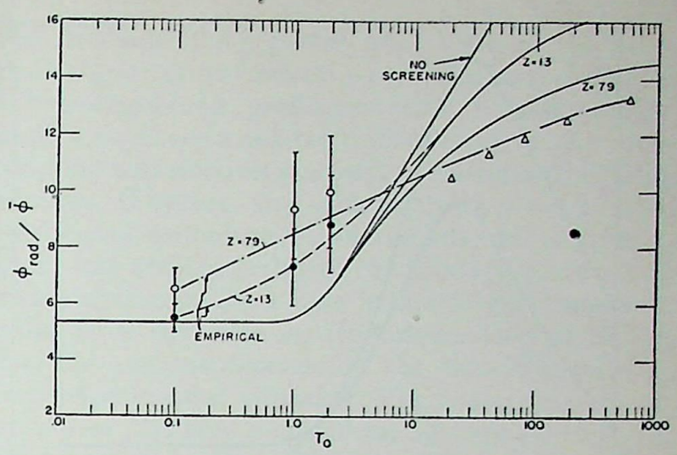


FIG. 22. Dependence of the total radiation cross section, $\phi_{\rm rad}[=(1/E_0)\int_0^{T_0}k d\sigma_k]$, on the initial electron kinetic energy, T_0. The solid lines are obtained from Formula 4BN for no screening, and from the numerical integration of Formulas 3BS(c) and 3BS(d) with screening corrections for Z equal to 13 and 79. The experimental points[33,35] are shown by the open and closed circles for a gold and aluminum target, respectively. The values shown by the triangles are estimated by numerical integration of Formula 3CS for $Z=79$. On the basis of the experimental data at the low energies and the theoretical values (triangles) predicted by the exact theory at the extreme relativistic energies, the dashed curves have been drawn as an estimate of the most accurate $\phi_{\rm rad}$ values for Z equal to 13 and 79.

B. Cross-Section Formulas for Free Electrons

(1) *Nonrelativistic Energies*

In contrast to the electron-nucleus and electron-positron systems, the electron-electron system has no

TABLE V. Corrected cross-section formulas for $d\sigma_k$.

Kinetic energy range for incident electron, Mev	Corrected cross-section formula[a]	Restrictions	Estimated accuracy[b]
0.01–0.10	$d\sigma_k=f_E d\sigma_k^{3BN(a)}$	$k>0.01T_0$	±5%
0.10–2.0	$d\sigma_k=Af_E d\sigma_k^{3BN}$	$k>0.01T_0$	±20%
2.0–15	$d\sigma_k=Ad\sigma_k^{3BN}$	$\gamma>15$	b
	$=Ad\sigma_k^{3BS(d)}$	$2<\gamma<15$	±5%
	$=Ad\sigma_k^{3BS(c)}$	$\gamma<2$	±5%
15–50	$d\sigma_k=d\sigma_k^{3BN}$	$\gamma>15$	b
	$=Ad\sigma_k^{3BS(d)}$	$2<\gamma<15$	±3%
	$=Ad\sigma_k^{3BS(c)}$	$\gamma<2$	±3%
50–500	$d\sigma_k=d\sigma_k^{3BN}$	$\gamma>15$	b
	$=d\sigma_k^{3CS(a)}$	$2<\gamma<15$	±3%
	$=d\sigma_k^{3CS(b)}$	$\gamma<2$	±3%

where f_E is defined in Formula (II-6), A is the correction factor given in Fig. 23, γ is equal to the quantity $100k(E_0EZ^{\frac{1}{2}})^{-1}$.

[a] The superscripts for $d\sigma_k$ give the formula numbers defined in Sec. IIC.
[b] No estimated accuracy is given at photon energies near the high-frequency limit of the spectrum. If better accuracy is desired in this region, the cross section at the high-frequency limit can be obtained from the dashed curves in Fig. 12, and the spectrum shape may be adjusted by fitting this end to the curves given by the formulas in column 2 above.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

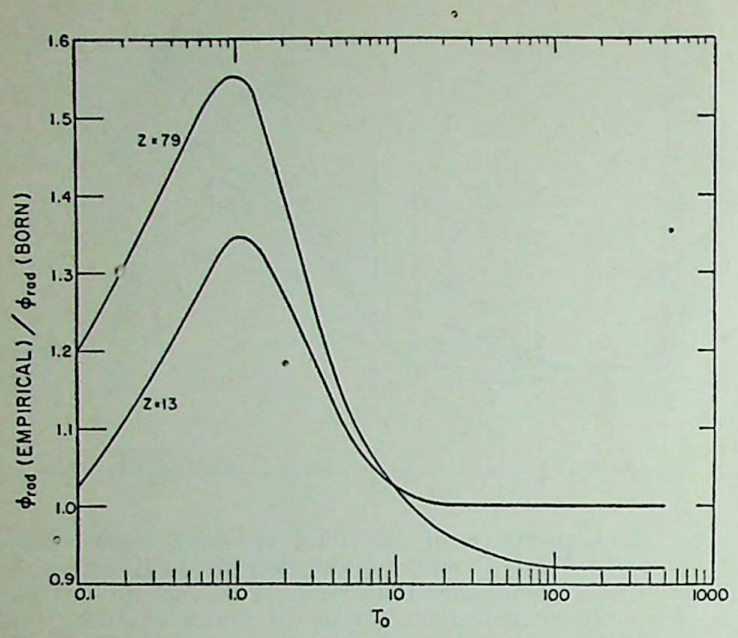


FIG. 23. Approximate correction factors for the Born-approximation ϕ_{rad} values with screening shown in Fig. 22. These factors have been estimated from the ratios of the empirical (dashed and dot-dashed) curves to the Born-approximation curves with screening in Fig. 22.

dipole moment. Therefore the electron-electron bremsstrahlung cross section becomes zero for calculations based only on the nonrelativistic dipole approximation. Garibyan[40] has made calculations beyond the dipole approximation and has obtained the following nonvanishing result for the cross-section differential in photon energy:

$$d\sigma_k'=\frac{r_0^2Z}{137}\left(\frac{8}{15}\frac{\beta}{\beta_0^3}\frac{dk}{k}\right)\left[17-\frac{3(\beta_0^2-\beta^2)^2}{(\beta_0^2+\beta^2)^2}+\left(\frac{\beta_0^2+\beta^2}{\beta_0\beta}+26\frac{\beta_0\beta}{\beta_0^2+\beta^2}-\frac{24\beta_0^3\beta^3}{(\beta_0^2+\beta^2)^3}\right)\times\ln\left(\frac{\beta_0+\beta}{\beta_0-\beta}\right)\right], \quad \text{(III-2)}$$

which for $k\rightarrow 0$ becomes[39]

$$d\sigma_{k\rightarrow 0}'=\frac{r_0^2Z}{137}\left[\frac{32}{5}\frac{dk}{k}\frac{1}{\beta_0^2}\left(\ln\frac{4\beta_0^2}{k}+\frac{17}{12}\right)\right]. \quad \text{(III-3)}$$

These results are only valid for $T_0\ll 1$.

TABLE VI. Maximum photon energy for electron-electron bremsstrahlung.

T_0	k_{max} $\theta_0=0°$	$\theta_0=90°$
100	99.	0.98
10	9.7	0.83
1.0	0.79	0.33
0.1	0.069	0.048
0.01	0.0054	0.0050

[40] G. M. Garibyan, Zhur. Eksptl. i Teoret. Fiz. 24, 617 (1953).

(2) *Extreme-Relativistic Energies*

Several calculations[39] based on the extreme-relativistic approximations give the following approximate formula for the cross-section differential in photon energy:

$$d\sigma_k'=\frac{4r_0^2Z}{137}\frac{dk}{k}\left[\left(1+\left(\frac{E}{E_0}\right)^2-\frac{2}{3}\frac{E}{E_0}\right)\times\left(\ln\frac{2E_0E}{k}-\frac{3}{2}\right)-\frac{E_0E}{9}\right], \quad \text{(III-4)}$$

which is similar in form to the electron-nuclear cross-section Formula 3BN(b).

The total radiation cross section obtained from Formula (III-4) is given as

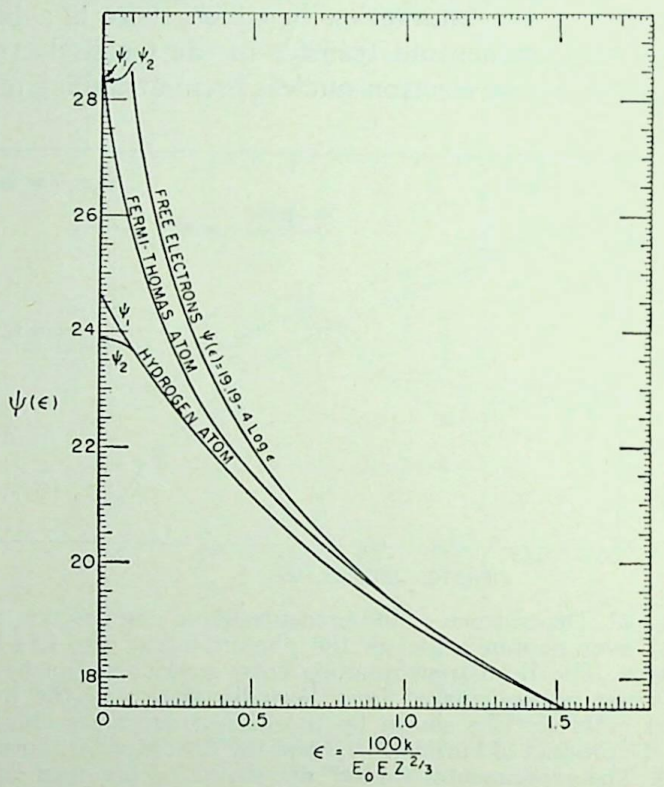


FIG. 24. Screening factors,[41] ψ_1 and ψ_2, for electron-electron bremsstrahlung. The curve marked "Hydrogen atom" was calculated[41] with exact wave functions. For free electrons, $\psi_1=\psi_2=\psi$.

$$\phi_{rad}'=\frac{4r_0^2Z}{137}[\ln(2E_0)-\tfrac{4}{3}]. \quad \text{(III-5)}$$

C. Cross-Section Formulas with Binding Corrections

The influence of atomic binding on the electron-electron bremsstrahlung cross section has been calculated only in the extreme-relativistic approximation. With the Thomas-Fermi model, the corrected formula for the cross-section differential in photon energy is[39]

$$d\sigma_k'=\frac{4r_0^2Z}{137}\frac{dk}{k}\left[\left(1+\left(\frac{E}{E_0}\right)^2\right)(\tfrac{1}{4}\psi_1(\epsilon)-1-\ln Z^{\frac{2}{3}})-\frac{2}{3}\frac{E}{E_0}(\tfrac{1}{4}\psi_2(\epsilon)-\tfrac{5}{6}-\ln Z^{\frac{2}{3}})\right], \quad \text{(III-6)}$$

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

where ϵ is equal to $100k(E_0EZ^{\frac{1}{3}})^{-1}$, and ψ_1 and ψ_2 are given[41] by the data in Fig. 24. For complete screening where $\epsilon \sim 0$, the cross section becomes

$$d\sigma_k' = \frac{4r_0^2Z}{137}\frac{dk}{k}\left(1+\left(\frac{E}{E_0}\right)^2-\frac{2}{3}\frac{E}{E_0}\right)\ln\frac{530}{Z^{\frac{1}{3}}}. \quad \text{(III-7)}$$

The total radiation cross section which is obtained for the complete screening case from (III-7) is given by

$$\phi_{\text{rad}}' = (4r_0^2Z/137)\ln(530/Z^{\frac{1}{3}}). \quad \text{(III-8)}$$

A comparison of this Formula (III-8) with the electron-nuclear bremsstrahlung cross-section Formula 4BS shows that the Z electrons in an atom increase the electron-nuclear cross section by the factor η so that the total cross section becomes

$$\phi_{\text{rad}}{}^{\text{total}} = Z(Z+\eta)(\phi_{\text{rad}}{}^{\text{4BS}}/Z^2). \quad \text{(III-9)}$$

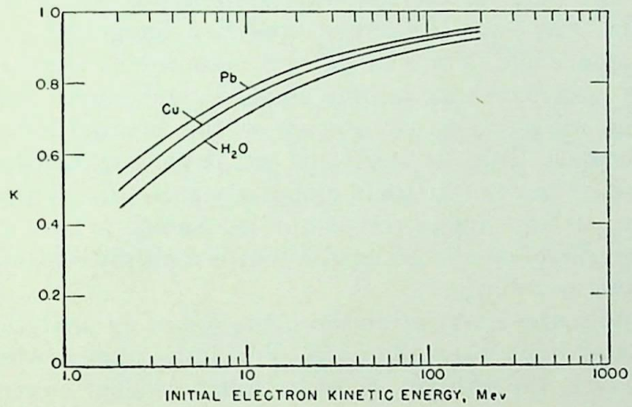


FIG. 25. Dependence of the radiation probability correction factor, $K(=\phi_{\text{rad}}/\phi_{\text{rad}}{}^*)$, on the initial electron kinetic energy and the target atomic number.

For complete screening, η is given by

$$\eta = \ln\frac{530}{Z^{\frac{1}{3}}}\Big/\left(\ln\frac{183}{Z^{\frac{1}{3}}}+\frac{1}{18}\right), \quad \text{(III-10)}$$

which varies from 1.04 for magnesium to 0.88 for lead. For most cases, a value of η equal to unity is sufficiently accurate.

IV. THICK-TARGET BREMSSTRAHLUNG PRODUCTION

Bremsstrahlung is produced in *thick* targets for most practical cases. In this discussion, a target is defined to be *thick* if the scattering and energy loss processes that occur as the electrons traverse the target have an appreciable influence on the bremsstrahlung production. In principle, a complete description of the bremsstrahlung emitted from a given target can be obtained from the cross sections for the pertinent elementary processes. For example, the angular distribution of the total bremsstrahlung power, the shape of the bremsstrahlung spectrum from an x-ray tube, or the efficiency of bremsstrahlung production can be calculated if detailed data are available with regard to the bremsstrahlung and electron scattering (elastic and inelastic) processes. However, any such analysis is necessarily a complicated procedure, since the calculations for the energy loss and scattering of the primary electrons and the absorption of the x-rays in the target must be included with the cross-section information of Sec. II. Also, the analysis depends on the characteristics of a given experimental situation. For these reasons, this paper does not give a complete, systematic treatment of thick-target bremsstrahlung production; instead it is confined to the presentation of pertinent experimental data as well as useful analytical results and procedures. Also, emphasis is placed on thick-target results that give absolute data on photon intensities and bremsstrahlung production efficiencies.

Some of the analytical results for thick-target bremsstrahlung are most conveniently expressed in terms of certain quantities which are defined in the following discussion. When an electron traverses a target, the average energy lost in the path length element dx by radiation can be written as

$$-dE_0 = NE_0(K\phi_{\text{rad}}{}^*)dx = KE_0dt, \quad \text{(IV-1)}$$

where N is the number of target atoms per cm^3 and $K\phi_{\text{rad}}{}^*$ is equal to the cross section ϕ_{rad} defined in Sec. IIC. $\phi_{\text{rad}}{}^*$ is equal to $(4Z^2r_0^2/137)\ln(183Z^{-\frac{1}{3}})$ cm^2, which is approximately the same as the expression for ϕ_{rad} at extreme-relativistic energies (see Formula 4BS). K is defined as the radiation probability correction factor and is plotted in Fig. 25 for various values of the target atomic number and the electron kinetic energy. The length t is given in units of the radiation length, t_0, which is defined as

$$t_0 = 1/N\phi_{\text{rad}}{}^* \text{ cm}. \quad \text{(IV-2)}$$

Values for t_0 in units of g/cm^2 as a function of the target atomic number are plotted in Fig. 26.

A. Thick-Target Bremsstrahlung Angular Distributions

(1) *Nonrelativistic and Intermediate Energies*[42]

For electron energies that are small or comparable to the electron rest energy, no analytical or empirical formulas have been derived for estimating the bremsstrahlung angular distribution from thick targets, and only a few experimental results are available.

In contrast to the extreme-relativistic region, the radiation intensity produced at these low energies is

[41] J. A. Wheeler and W. E. Lamb, Phys. Rev. **55**, 858 (1939), and **101**, 1836 (1956).

[42] The results that are presented for the nonrelativistic and intermediate energy region where $T_0 \leq 1$ apply only to targets that are thick enough to stop the electrons. For the relativistic region where $T_0 \gg 1$, there is no such restriction on the target thickness.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

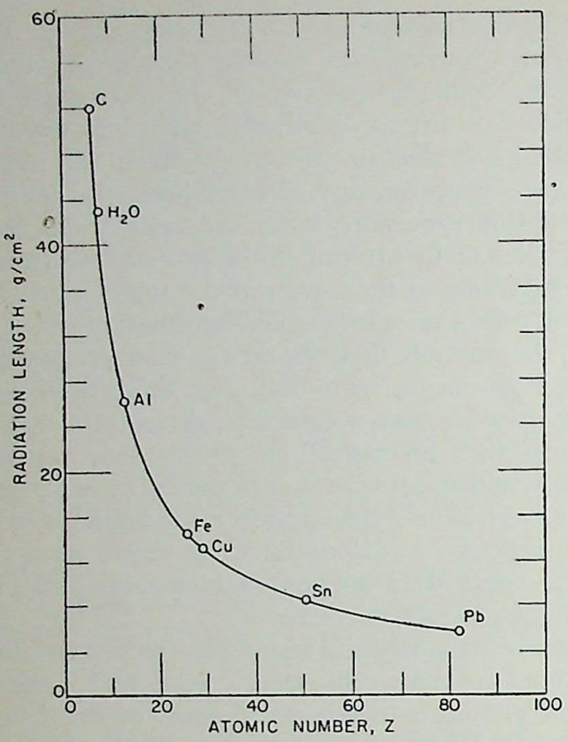


FIG. 26. Radiation lengths in grams per square centimeter for various materials.

important at large angles, and is about the same order of magnitude at both zero and ninety degrees. However, because the absorption of the bremsstrahlung photons in the target is large, the angular distribution of the bremsstrahlung is largely dependent on the target geometry in specific experimental situations. One of the few examples in which angular distribution data are presented in a more general way with corrections for the geometry and the target absorption is to be found in the measurements of Buechner, Van de Graaff, Burrill, and Sperduto[43] for initial electron kinetic energies in the region from 1.25 to 2.35 Mev. Their results for beryllium and gold targets are given in Fig. 27. The curves show the angular dependence of the radiation intensity integrated over photon energy for specified electron energies. These data indicate that the intensity ratio at zero and ninety degrees is approximately 10 for beryllium and 3 for gold at 1.5 Mev, and approximately 40 for beryllium and 4 for gold at 2.35 Mev. Also from these data, we can obtain the following empirical expressions for the power radiated at zero degrees:

$$I_{(\alpha=0)}(\mathrm{Au}) = 9.4(\mathcal{T}_0)^{2.9} \text{ roentgens per minute per ma at 1 meter for gold,}$$
$$I_{(\alpha=0)}(\mathrm{Be}) = 0.92(\mathcal{T}_0)^{3.4} \text{ roentgens per minute per ma at 1 meter for beryllium,} \quad \text{(IV-3)}$$

where $\mathcal{T}_0$ is the electron kinetic energy in m_0c^2 units for the electrons incident on the target, and α is the angle between the photon direction and the direction of the incident electron beam. Then with the approximate conversion factor of one roentgen equal to 3000 ± 500 ergs/cm^2 for photons with energies in the range from 0.1 to 2 Mev,[44] we have

$$I_{(\alpha=0)}(\mathrm{Au}) = 0.5(\mathcal{T}_0)^{2.9} \text{ watts/ma-steradian,}$$
$$I_{(\alpha=0)}(\mathrm{Be}) = 0.05(\mathcal{T}_0)^{3.4} \text{ watts/ma-steradian.} \quad \text{(IV-4)}$$

From these equations, the fraction, R, of the total incident electron kinetic energy that is radiated per steradian at zero degrees is

$$R_{(\alpha=0)}(\mathrm{Au}) = 10^{-3}(\mathcal{T}_0)^{1.9} \text{ for gold,}$$
and
$$R_{(\alpha=0)}(\mathrm{Be}) = 10^{-4}(\mathcal{T}_0)^{2.4} \text{ for beryllium.} \quad \text{(IV-5)}$$

(2) *Relativistic Energies*

At high energies, estimates of the bremsstrahlung angular distribution from thick targets have been made on the basis of the following simplifying approximations. First, the thin-target spectrum integrated over photon angle (Formula 3BS) is assumed to represent the spectrum shape for any angle. Second, the intrinsic (thin target) angular spread of the bremsstrahlung (Formula 2BS) is neglected at large angles where $\alpha \gg E_0^{-1}$ but not at small angles where $\alpha \leq E_0^{-1}$; therefore, at large angles the photon is assumed to have the same direction as the electron that is multiply scattered before it radiates.

With these approximations, the following analytical results have been obtained. For large angles where $\alpha \gg E_0^{-1}$, the fraction, R, of the total incident electron

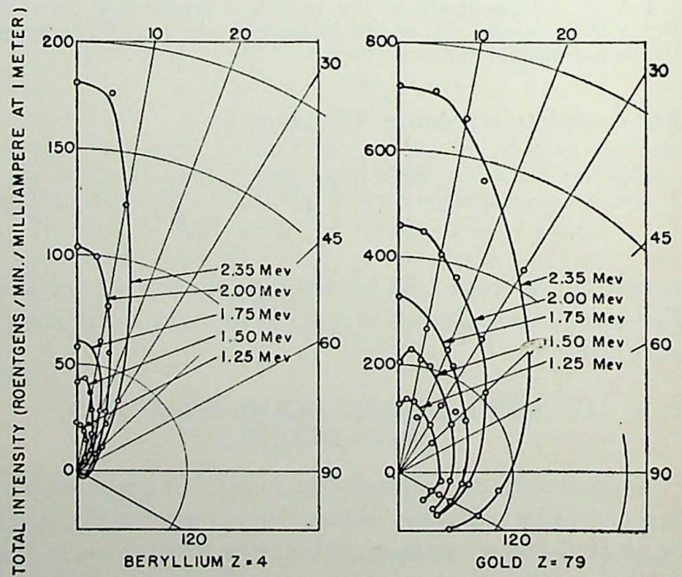


FIG. 27. Angular dependence of the thick-target bremsstrahlung intensity integrated over photon energy for 1.25- to 2.35-Mev electrons. These results were obtained by Buechner, Van de Graaff, Burrill, and Sperduto[43] and include corrections for the target.

[43] Buechner, Van de Graaff, Burrill, and Sperduto, Phys. Rev. 74, 1348 (1948).

[44] W. V. Mayneord, Brit. J. Radiol. Suppl. No. 2, 136 (1950). For photons outside this energy range, the conversion factor has a significant energy dependence and the factor must be weighted by the bremsstrahlung spectrum shape.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

kinetic energy that is radiated per steradian at the angle, α, is given[45] as

$$R_{(\alpha \gg E_0^{-1})} = \frac{K\mathcal{E}_0^2}{1760\pi} \mathrm{Ei}\left[\frac{-\mathcal{E}_0^2\alpha^2}{1760t}\right], \qquad \text{(IV-6)}$$

where $\mathcal{E}_0$ is the total energy in m_0c^2 units of the electron incident on the target, and Ei is the exponential integral[46]

$$-\mathrm{Ei}(-y) = \int_y^\infty \frac{e^{-z}}{z} dz, \quad >0 \text{ for } y>0. \qquad \text{(IV-7)}$$

For small angles where $\alpha \lesssim E_0^{-1}$, Muirhead, Spicer, and Lichtblau[47] have obtained the following expression for the bremsstrahlung angular distribution

$$R_{(\alpha \lesssim E_0^{-1})} = \frac{K\mathcal{E}_0^2}{1760\pi}\left\{-\mathrm{Ei}\left[\frac{-\mathcal{E}_0^2\alpha^2}{1760t}\right] + \mathrm{Ei}\left[\frac{-\mathcal{E}_0^2\alpha^2}{7.15}\right]\right\}. \qquad \text{(IV-8)}$$

This formula gives good agreement with experimental data[47] and can readily be evaluated at small angles by keeping the first term in the expansion of the exponential integrals which is

$$\mathrm{Ei}(-z_1) - \mathrm{Ei}(-z_2) \rightarrow \ln(z_1/z_2) \quad \text{for} \quad z_1, z_2 \rightarrow 0. \qquad \text{(IV-9)}$$

Thus

$$R_{(\alpha=0)} = \frac{K\mathcal{E}_0^2}{1760\pi} \ln 246t, \text{ for } t \gg 2\times10^{-3}. \qquad \text{(IV-10)}$$

For $t=0.1$ and $\mathcal{E}_0=3$, this formula agrees reasonably well with the result predicted by the low-energy Formula (IV-5). For thin targets, this "on-axis" intensity becomes[45]

$$R_{(\alpha=0)} = Kt\mathcal{E}_0^2/4\pi, \quad \text{for} \quad t \ll 2\times10^{-3}. \qquad \text{(IV-11)}$$

Estimates of the ratio $R_\alpha/R_{(\alpha=0)}$ for tungsten ($Z=74$) are given in Fig. 28 for three target thicknesses.

Several conclusions for high-energy angular distributions can be drawn from the form of the Formulas (IV-8), (IV-10), and (IV-11). The logarithmic form of Formula (IV-10) shows that most of the radiation comes from the front part of the target. Also, since the fractional energy radiated depends on $\mathcal{E}_0^2$, the total energy radiated at zero degrees will depend on $\mathcal{E}_0^3$. Two additional effects influence the dependence on $\mathcal{E}_0$ of the total radiated energy. The factor K, according to Fig. 25, increases slightly with $\mathcal{E}_0$, and for very thick targets the effective t in Formula (IV-10) will increase logarithmically with $\mathcal{E}_0$. Therefore, the total energy radiated on the axis of the bremsstrahlung beam will depend on at least a 3.2 exponent for a thin target and on a slightly higher exponent for a thick target. The specific exponent to be used will depend on the energy range of interest, the effective target thickness, and the experimental geometry.

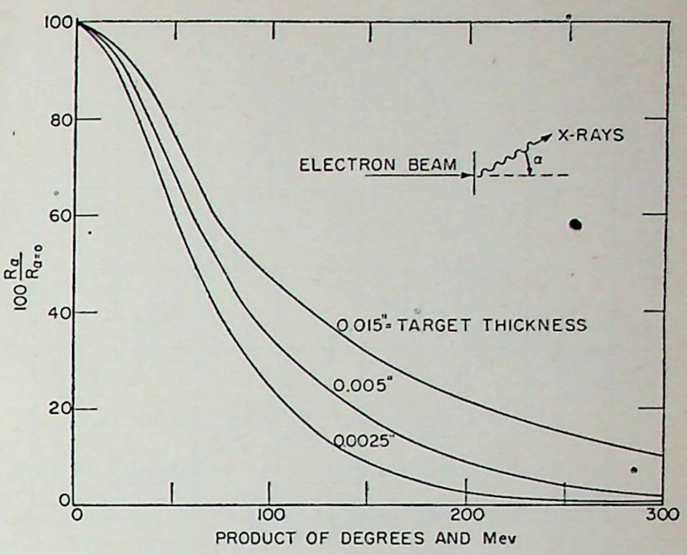


FIG. 28. Theoretical bremsstrahlung angular distributions from thick tungsten targets for relativistic energies. These data are obtained from the National Bureau of Standards Handbook 55. R_α is defined as the fraction of the total incident electron kinetic energy that is radiated per steradian at the angle α.

B. Thick-Target Bremsstrahlung Spectra

(1) *Nonrelativistic and Intermediate Energies*[42]

In this low-energy region the radiation has a broad angular distribution (see IVA), and the dependence of the spectrum shape on photon angle is important.[48] No general analytical expressions which accurately predict the spectrum as a function of angle for any experimental situation are available at these energies. Part of the difficulty has been the inadequacy of the Born-approximation cross-section differential in photon energy and angle (Formula 2BN). Nevertheless it has been possible to obtain reasonable agreement between theoretical and experimental thick-target spectrum shapes shown in Fig. 29 for a particular application[49] with an initial electron kinetic energy of 1.4 Mev, photon angles of zero and ninety degrees, and a tungsten target. In this example, the experimental results confirm the theoretical dependence of the spectrum shape on photon angle after distortions due to photon absorption in the target and surrounding materials are eliminated. The results also show that the relative number of photons in the high-frequency region increases as the emission angle becomes smaller. This trend is just opposite to the behavior observed for thin-target spectra.[35]

[45] J. D. Lawson, Nucleonics **10**, No. 11, 61 (1952).
[46] Exponential integral functions are tabulated in *National Bureau of Standards Tables of Sine, Cosine, and Exponential Integrals* (U. S. Government Printing Office, Washington, D. C.), Vols. 1 and 2 (MT5 and MT6).
[47] Muirhead, Spicer, and Lichtblau, Proc. Phys. Soc. (London) **A65**, 59 (1952).
[48] One of the earliest experimental indications of this dependence was found by C. E. Wagner, Physik. Z. **21**, 621 (1920).
[49] Miller, Motz, and Cialella, Phys. Rev. **96**, 1344 (1954).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

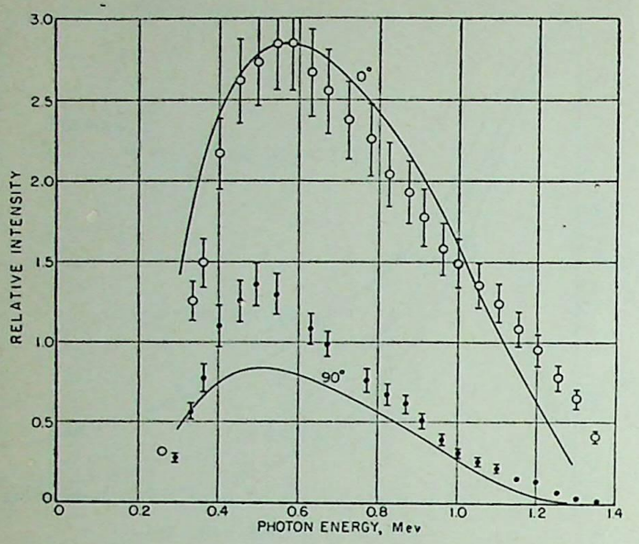


FIG. 29. Relative spectral intensities at 0° and 90° for 1.4-Mev electrons incident on a thick-tungsten target.[49] The solid curves are obtained from theoretical estimates that include electron scattering effects and photon absorption in the materials surrounding the target. The experimental values have been normalized and are shown by the open (zero degrees) and closed (90 degrees) circles. To obtain absolute spectral intensities in Mev per steradian per Mev per incident electron, the ordinate should be multiplied by 10^{-3} for the theoretical curves and by 2.1×10^{-3} for the experimental points.

With regard to estimates of the shape of the spectrum integrated over the photon direction, Kramers[50] obtained the following simple, analytical expression:

$$I_k = AZ(k_0-k), \qquad \text{(IV-12)}$$

where I_k is the energy radiated in all directions in the energy interval, $(k, k+dk)$, A is a proportionality constant, and k_0 is the photon energy at the high-frequency limit. This result was derived on the basis of a nonrelativistic, semiclassical calculation, in which electron scattering effects (including backscattering) were neglected and only the electron energy loss was considered. In spite of these limitations and because of its simplicity, the Kramers Formula (IV-12) has been used extensively to estimate the thick-target spectrum (not including the characteristic radiation[51]) at a given angle for various experimental cases, with corrections included for the photon absorption in the target and surrounding materials. Results obtained for various electron energies in this low-energy region have shown general qualitative agreement between the theoretical (Kramers) and experimental spectrum shapes,[52] and indicate that Formula (IV-12) is satisfactory, at least for order of magnitude estimates.

(2) *Relativistic Energies*[42]

Two complementary procedures for calculating thick-target spectra at high energies which include effects of electron scattering in the target are given by Penfold[53] and Hisdal.[54] The Penfold calculations estimate the thick-target effects primarily in the high-frequency region and give the spectrum integrated over photon directions up to a maximum angle, Γ, with respect to the direction of the incident electron beam. The Hisdal calculations estimate the thick-target effects on the over-all spectrum shape in the forward direction and should not be applied to the high-frequency region.

The Penfold method[53,55] assumes that (a) the Schiff Formula 3BS(e) describes the intrinsic spectrum at all angles; (b) the electron energy loss rather than electron scattering in the target produces the predominant effect on the shape of the spectrum for large values of Γ; (c) no electron radiates more than one photon; and (d) the photon absorption in the target is negligible. With these approximations, Penfold obtained the following formula for the thick-target spectrum integrated over photon direction to a maximum angle Γ determined by the detector:

$$P_k = nN_0\int_{k+1}^{\mathcal{E}_0} S(\mathcal{E}_0,E_0,\Gamma,x_0)\,d\sigma_k\,dE_0, \qquad \text{(IV-13)}$$

where P_k is the number of photons in the energy interval k to $k+dk$, N_0 is the number of target atoms per cm^3, n is the number of electrons incident on the target, $\mathcal{E}_0$ is the total energy of the incident electron in m_0c^2 units, x_0 is the target thickness in g/cm^2, and $d\sigma_k$ is given by Formula 3BS(e) for electrons with energy E_0. The function S represents the probability that radiation produced by the electrons reaches the detector, and can be written as

$$S(\mathcal{E}_0,E_0,\Gamma,x_0) = \int_0^{x_0} B_3(\mathcal{E}_0,E_0,\Gamma,x)B_4(\Gamma,E_0,x)B_5(x_0,x)\,dx, \qquad \text{(IV-14)}$$

where the function B_3 gives the fraction of the radiation emitted by electrons with energy E_0 at the target depth x, B_4 is the fraction of electrons that penetrate beyond the thickness x, and B_5 accounts for path length straggling. These B functions require involved numerical evaluations, and the results are described in detail in the Penfold report.[53] Motz, Miller, and Wyckoff[56] have estimated the thick-target spectrum for a particular experimental situation in which the brems-

[50] H. A. Kramers, Phil. Mag. **46**, 836 (1923).
[51] For estimates of the thick-target characteristic radiation intensities that are superimposed on the continuous spectrum, see A. H. Compton and S. K. Allison, *X-Rays in Theory and Practice* (D. Van Nostrand Company, Inc., Princeton, New Jersey, 1949), pp. 69–89.
[52] A detailed summary of such comparisons is given in Natl. Bur. Standards (U.S.) Handbook **62**, 20–24 (1957).
[53] A. Penfold, University of Illinois Report (unpublished).
[54] E. Hisdal, Phys. Rev. **105**, 1821 (1957); E. Hisdal, Arch. for Math. Naturvidenskab **54**, No. 3, 1 (1957).
[55] A similar but less general method has been used by R. Wilson, Proc. Phys. Soc. (London) **A66**, 038 (1953). Wilson's calculations did not include electron scattering effects in the target and his results give a spectrum shape averaged over the photon directions.
[56] Motz, Miller, and Wyckoff, Phys. Rev. **89**, 968 (1953).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

strahlung is produced with an 11.3 Mev (kinetic energy) electron beam incident on a tungsten target (approximately 0.010-inch thick), and is measured on the beam axis with a small detector ($\Gamma\sim 0$). They used the following simplified, analytical form for the thick-target generating function:

$$S=1-\exp\left\{-\frac{1}{51(\mathcal{E}_0-E_0)}\left(\frac{\mathcal{E}_0}{E_0}\right)^2\right\} \qquad \text{(IV-15)}$$

and, as shown in Fig. 30, have obtained good agreement with experimental results. For the more general calculations, Penfold used Formulas (IV-13) and (IV-14) to estimate the thick-target spectrum shapes for an incident electron kinetic energy of 15 Mev, a 0.020-inch platinum target, and for two detectors which subtend different angles on the electron beam axis, ($\Gamma=10$ degrees, $\Gamma\gg 10$ degrees). A comparison is made in Fig. 31 of these two Penfold results (curves C and D) with the spectrum shape predicted by Formula 3BS(e) (curve A) and with the shape resulting from the application of the S function in Formula (IV-15) (curve B). The curves show that Formulas 3BS(e) and (IV-15) give a greater number of photons in the high-frequency region relative to the total number in the spectrum compared with the more accurate spectral shape predicted by the Penfold procedure. For certain cases, the spectrum shape predicted by the simplified Formula (IV-15) may be sufficiently accurate.

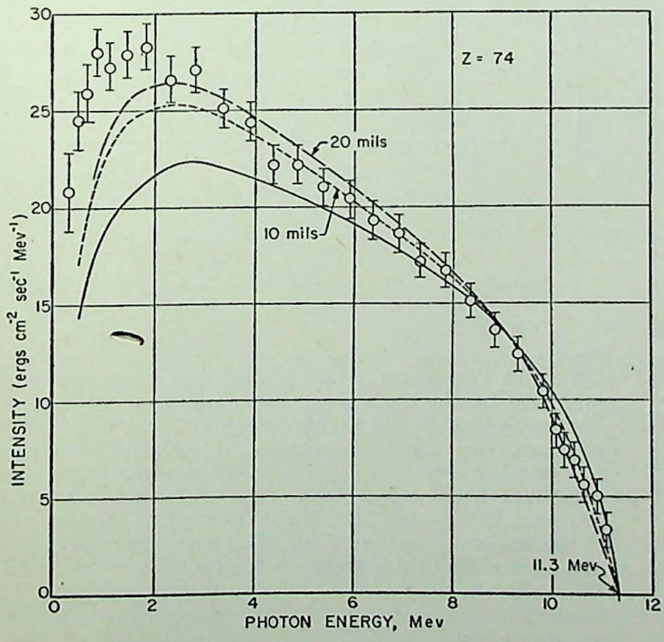


FIG. 30. Bremsstrahlung intensity spectrum in the forward direction for 11-Mev electrons incident on a thick-tungsten target.[56] The thin-target Born spectrum, modified by the photon absorption in the materials surrounding the target, is shown by the solid curve. The dashed curves show the spectra expected for a 10-mil and a 20-mil target, and the experimental values are given by the open circles.

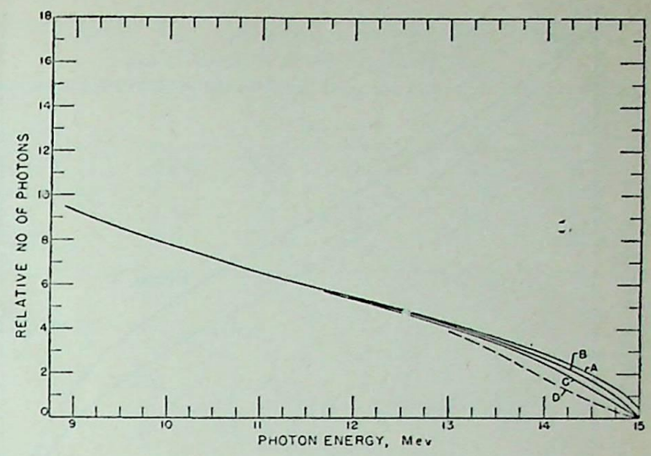


FIG. 31. Comparison of the spectrum shapes predicted by the thick target Penfold calculations[53] and by Schiff's thin target Formula 3BS(e) for 15-Mev electrons. The Schiff curve A shows the spectrum integrated over photon angle. Curve B is the spectrum shape predicted by Formula (IV-13) with the simplified S function given by Formula (IV-15). Curve C is the spectrum shape predicted by the Penfold calculations [Formulas (IV-13) and (IV-14)] for a 0.020-in. platinum target and a detector which subtends an angle, Γ, of 10 degrees. Curve D gives the Penfold spectrum for $\Gamma\gg 10$ degrees.

The Hisdal method[57] assumes that (a) the spectrum variation with angle as given by the Schiff Formula 2BS has the dominant effect on the thick-target spectrum shape; (b) the energy loss of electrons in the target is negligible; (c) no electron radiates more than one photon; and (d) the photon absorption in the target is negligible. With these approximations, Hisdal has calculated tables for estimating the spectrum shape to be expected in a small detector placed on the electron beam-target axis. Hisdal's results are given in terms of a correction factor which multiplies Formula 3BS(e) for a given value of k to obtain the corrected spectrum for a particular target thickness. Examples of spectra calculated by Hisdal's method are given in Figs. 32(a)–(e) for 10-, 20-, 40-, 90-, and 300-Mev electrons, and are compared with the Schiff spectra integrated over photon direction [Formula 3BS(e)]. When the detector subtends a large solid angle at the target, the measured spectrum shape is given by the cross section integrated over the photon direction. Figures 7 and 8 give data for the spectrum shape included within a given detector angle. If this shape for a given experimental arrangement is estimated to be similar to the zero degree spectrum, then the Hisdal correction will be important; if this shape is estimated to be more similar to the spectrum integrated over all angles, then Hisdal's correction will be unimportant.

C. Efficiency for Bremsstrahlung Production

The bremsstrahlung production efficiency for a given electron kinetic energy and target material is defined

[57] Similar calculations that apply only to a specific target material and electron energy have been made by A. Sirlin, Phys. Rev. **106**, 637 (1957). While the Hisdal calculations include only the Schiff complete screening approximations, Sirlin examined the effects of both the intermediate and complete screening approximations on the spectrum shape.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

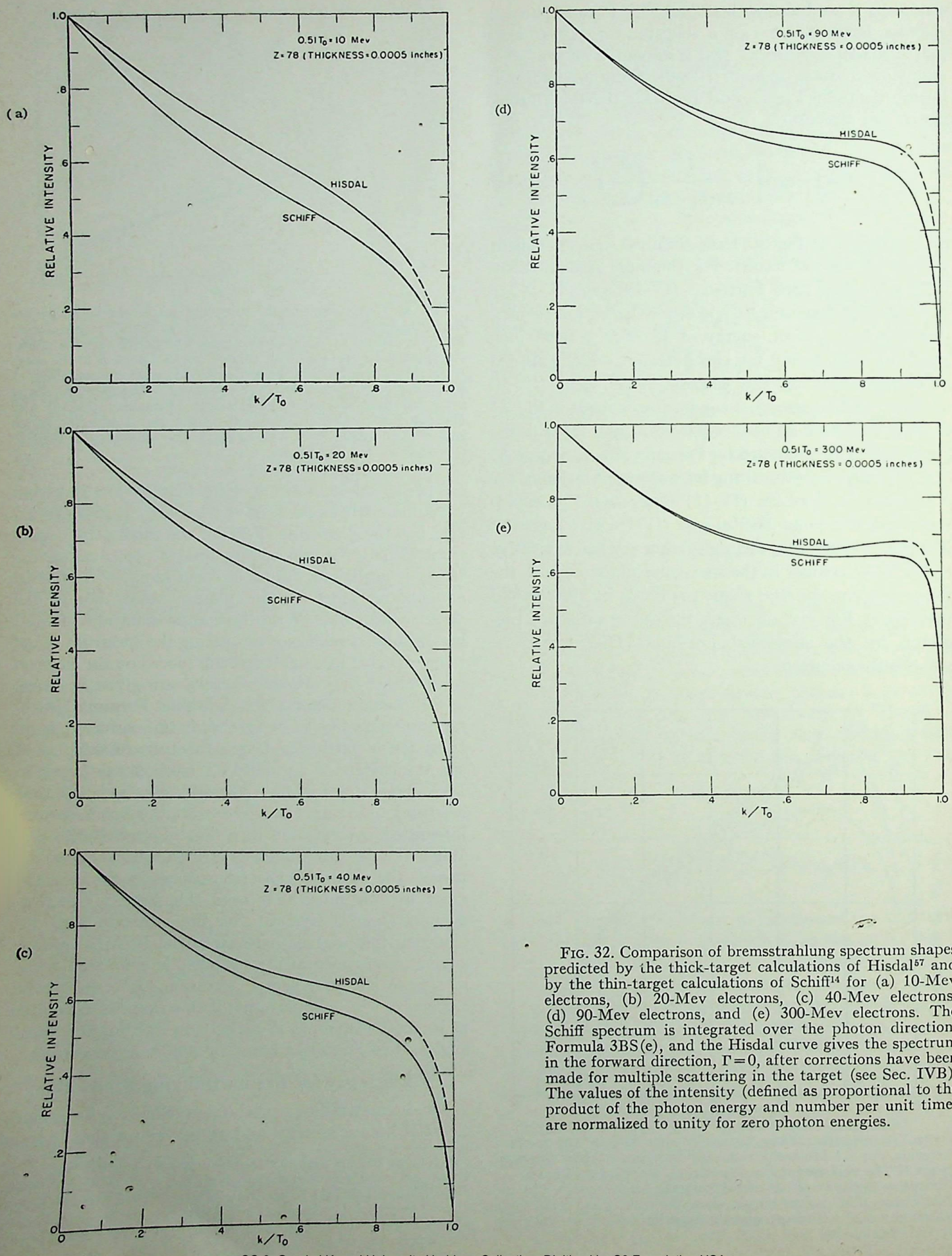


FIG. 32. Comparison of bremsstrahlung spectrum shapes predicted by the thick-target calculations of Hisdal[67] and by the thin-target calculations of Schiff[14] for (a) 10-Mev electrons, (b) 20-Mev electrons, (c) 40-Mev electrons, (d) 90-Mev electrons, and (e) 300-Mev electrons. The Schiff spectrum is integrated over the photon direction, Formula 3BS(e), and the Hisdal curve gives the spectrum in the forward direction, $\Gamma=0$, after corrections have been made for multiple scattering in the target (see Sec. IVB). The values of the intensity (defined as proportional to the product of the photon energy and number per unit time) are normalized to unity for zero photon energies.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

as the ratio of the total bremsstrahlung power radiated when an electron current is incident on a target to the total power in the incident electron beam. The results of various theoretical and experimental determinations of the efficiency are given in the following.

(1) *Nonrelativistic and Intermediate Energies*

The efficiency results for this energy region apply only to the cases in which the electrons expend all of their kinetic energy in the target. Experimental determinations are complicated by (a) the large photon absorption in the target and (b) the large electron backscattering from the target.[49] In the available measurements of the efficiency, corrections have been made for effect (a) but not for effect (b). Therefore, these experimental efficiencies are less than the values that would be obtained if all of the electrons were completely stopped in the target.

In this low-energy region, most experimental and theoretical results[58] are in agreement within a factor of two with an efficiency, ϵ, given by the following formula:

$$\epsilon = 5\times 10^{-4} Z \mathcal{T}_0. \qquad \text{(IV-16)}$$

(2) *Relativistic Energies*

(*a*) *Intermediate thickness targets.*—A target is defined to have an intermediate thickness if the incident electrons do not expend all of their energy as they traverse the target. This condition usually exists in high-energy electron accelerators.

The efficiency of bremsstrahlung production for targets having an intermediate or small ($\ll t_0$) thickness can be estimated from the expression

$$\epsilon = [(d\mathcal{E}_0)_R/\mathcal{T}_0] \sim Kt, \qquad \text{(IV-17)}$$

where $(d\mathcal{E}_0)_R$ is the energy loss by radiation, t is the target thickness in units of the radiation length t_0 [Formula (IV-2)] and K is the radiation probability correction factor given in Fig. 25.

(*b*) *Thick targets.*—For thick targets, the incident electrons lose all of their energy in the target. Formula (IV-16) obviously does not apply at high energies for which the efficiency must remain less than one. An approximate relation for the efficiency in this high energy region has been derived[59] by assuming that the total electron energy loss per unit path length can be written as

$$-\frac{d\mathcal{E}_0}{dx} = -\left(\frac{\rho Z}{A}\right)\{6+3.5\times 10^{-3}\mathcal{E}_0 Z\}, \qquad \text{(IV-18)}$$

where ρ is the density and A the atomic weight of the target material, and where the first term is the collision loss and the second term is the radiation loss. Integration from the initial energy $\mathcal{E}_0$ to 1 gives the following distance, x_0, traveled by an electron in losing all of its energy:

$$x_0 = \frac{A}{3.6\times 10^{-3}\rho Z^2}\ln(1+6\times 10^{-4} Z\mathcal{E}_0). \qquad \text{(IV-19)}$$

Then the efficiency becomes

$$\epsilon = 1-\frac{\text{collision loss}}{\mathcal{E}_0} = 1-\frac{6\rho Z}{A}\frac{x_0}{\mathcal{E}_0} = \frac{3\times 10^{-4} Z\mathcal{T}_0}{1+3\times 10^{-4} Z\mathcal{T}_0}. \qquad \text{(IV-20)}$$

This procedure does not account for the large fluctuations in the radiation loss process and, therefore, provides only a rough estimate of the efficiency. At low energies, Formula (IV-20) reduces to $3\times 10^{-4} Z\mathcal{T}_0$, which agrees roughly with Formula (IV-16). Some representative values of the efficiency obtained from Formula (IV-20) are given in Table VII. These values do not include corrections for the x-ray absorption in the target, which cannot be neglected for most situations.

TABLE VII. Approximate percentage efficiencies for x-ray production.

$\mathcal{T}_0$:	2	4	10	20	50	600
Carbon	0.36	0.72	1.77	3.47	8.3	52
Aluminum	0.77	1.54	3.75	7.2	16.3	70
Iron	1.54	3.0	7.2	13.5	28.1	82
Lead	4.7	9.0	19.7	33	55	94

V. ACKNOWLEDGMENTS

This report has benefitted considerably from discussions with Dr. U. Fano, Dr. M. Danos, Dr. J. Leiss, and Dr. H. O. Wyckoff. We are grateful to them for their helpful comments and criticisms. We express our special thanks to Dr. L. Maximon who critically reviewed the manuscript and made available a paper by Olsen and Maximon prior to publication. We are also indebted to Mr. John Hubbell for the comparison of the Hisdal and Sirlin results and for most of the computations and detailed plotting in the figures.

[58] These results are summarized by H. Kulenkampff, "Physics of the Electron Shells," Fiat Rev. Ger. Sci. **1939–1946**, 95; R. D. Evans, *The Atomic Nucleus* (McGraw-Hill Book Company, Inc., New York, 1955), p. 616; and by S. T. Stephenson (reference 1).

[59] H. W. Koch and J. W. Wyckoff, IRE Trans. on Nuclear Sci. NS-5, No. 3 (1958).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

Interaction of Normal Modes with Electron Traps

JORDAN J. MARKHAM

Research Department, Zenith Radio Corporation, Chicago, Illinois

1. Introduction 956
Part I. The Hamiltonian 958
2. Physical Concepts Involved in Electron-Phonon Interactions 958
3. Formal Theory 959
(a) Ground State 961
(b) Upper State 961
4. Physical Interpretation of Optical and Thermal Transformations 963
5. Properties of the Total Hamiltonian 964
(a) Trapped Electron 965
(b) Free Electron 965
(c) Comparison with Quantum Electrodynamics 966
6. Conclusions of Part I 967
Part II. Broadening of Absorption Lines Due to Multiphonon Processes 967
7. Introduction to Part II 967
8. General Formulation of Absorption Problem 969
9. Shape of the Band 972
(a) Low Temperature 974
(b) High Temperature 974
10. Method of Moments 977
(a) CC Model 978
(b) Huang-Rhys Problem 979
(c) Double-Frequency Model 979
11. Analysis of Experimental Data 980
12. Conclusions of Part II 982
13. Appendix I. List of Symbols 983
14. Appendix II. Meyer-Pekar Method 984
15. Appendix III. Mehler's Formula 985
16. Appendix IV. Pekarian Curve 985
17. Appendix V. Effects Which Occur When f is Not a Constant 986
(a) Oscillator Strength Depends on the Frequency 987
(b) Oscillator Strength Depends on the q's 987
Bibliography 988

1. INTRODUCTION

CONSIDER an electron trap in a solid with two levels, u (upper) and g (ground). A fundamental question is: how does an electron transition affect the normal vibrational modes of the crystal? The answer so far has been obtained either with the use of a phenomenological macroscopic theory, or by the use of configurational coordinate curves.

The phenomenological theory, due to Fröhlich (F1,* F2, and H4; also Born and Huang, B2, p. 82), ignores details of the lattice structure. In a polar crystal Fröhlich assumes that the forces on the ions, as well as the polarization, depend exclusively on the macroscopic electric field and relative ionic displacements at the point in question. This model leads to pure transverse and longitudinal modes. It is only true for the longest "optical" waves in which the positive and negative ions vibrate 180° out of phase. This theory does not account for the change in the character of the normal mode with decreasing wavelength. In NaCl the longitudinal optical branch in certain directions has a frequency drop of 57%, and the shortest waves are pure acoustical, in which the ions vibrate in phase (M3). Fröhlich's model is useful in the theory of reststrahlen where one deals with electromagnetic radiations whose wavelength is large compared to the inter-ionic distances, but this approach is questionable for phenomena which depend critically on short range effects.

The configurational coordinate model assumes that a very small number of local modes surround a point imperfection. This approach is supported by recent calculations. (See M2 and references therein.) One would expect the local modes to have frequencies in the gap between the acoustical and optical branches. This suggests that their frequency spread may be small. We refer to this model by the letters LM. It is assumed that there are one or several "effective" frequencies associated with the local modes. The extreme situation where there is only one mode is referred to by the letters CC (configuration coordinate, D1). The LM models cannot give a completely correct picture since long range effects must exist in polar solids. The Fröhlich model cannot either, since it ignores local lattice distortions.

One of the most important effects of the electron-phonon interaction is the broadening of optical absorption lines. Studies of these phenomena will give information regarding the effects of electronic transitions on normal modes.

In general, rigorous quantum mechanics has been used on this problem when the phenomenological model is employed. The original techniques required infinitely small interaction between the trapped electron and the modes. It, therefore, could not be used in the LM or CC models.

While these developments were taking place, Williams (W2), as well as Williams and Hebb (W3), developed

* Refer to items in the Bibliography at the end of the paper.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

an independent treatment which was applied to the CC model. This approach makes a series of approximations and is sometimes referred to as "semiclassical." A more detailed justification of Williams' method has been made by Lax (L1), Klick (K2), and Dexter (D1). This development shows that at very low temperatures the shape of an absorption band should be Gaussian (provided that the electronic transition does not induce a frequency change in the CC mode). Further, the position of maximum absorption is temperature independent even if the frequency changes.

Since 1950 the rigorous quantum mechanical method has been fully developed, and one can now apply it to various LM models. Williams' simpler approach is still very useful, since it gives a vivid picture of the absorption process. However, it is beset by many approximations, and the final results do not always agree with more rigorous calculations.

To summarize the development so far, we have several models—phenomenological (Fröhlich), local modes, and configurational coordinate. We have two means of calculating shapes of bands—Williams', and the rigorous quantum mechanical approach. For some models the second method can be carried through in detail. Most realistic ones, however, require Williams' approach. A rather formal theory for all models has been worked out by Kubo and Toyozawa (K6).

This paper attempts to synthesize the rigorous theory and to apply it to various models. We start by examining the Hamiltonian to be used (Part I). Fortunately, it can be written in a form that applies to all the models. We next discuss the properties of these Hamiltonians and point out the essential difference between phonon-electron and photon-electron interactions. Only in special types of solid state problems may they be considered similar. At this stage, one can see why multiphonon processes occur. Some of the results of Part I are known, but their interrelations have not been fully understood. This has led to appreciable confusion. Some of the mathematical techniques are those used by Born and Huang in connection with another problem.

After this development, the photon absorption problem is considered rigorously and in some detail (Part II). The Hamiltonian developed permits a unification of the field which is most desirable, since experimental results are being obtained.

Applications of the theory are not considered because the author does not trust the presently available data. The absorption band associated with the F center in KCl has been measured a great many times. It seems to fulfill the requirement stated in the first paragraph. Various laboratories, however, do not report the same results. A typical set of half-width, H, measurements are shown in Table I.†

At low temperatures H can be measured to about five parts per thousand so that the variations cannot be attributed to experimental error (M1). The author doubts that the scatter is related to the sample purity, and the temperature variation of H is too small to explain the discrepancies. Experiments have shown (K3) that, unless special care is taken, any values between 0.19 ev and 0.25 ev can be obtained for H (in KCl at 78°K). Improper quenching and very slight exposure to light produce new bands, and H then corresponds to a composite.

TABLE I. H of KCl.

θ (°K)	H (ev)	Observer
28	0.19	Mollow (M10)
4	0.18	Russell and Klick (R1)
4	0.17	Burstein and Oberly (B5)
5	0.168	Compton and Klick (C3)
5	0.17	Duerig and Markham (D2)
5	0.163	Markham and Konitzer (M4)
85 (App)	0.26	Molnar (M11)
87	0.23	Mollwo (M10)
87	0.243	Kanzaki (K1)
77	0.3	van Doorn and Haven (D3)
77	0.22	Russell and Klick (R1)
77	0.20	Burstein and Oberly (B5)
78	0.20	Duerig and Markham (D2)
78	0.193	Konitzer and Markham (K4)
77	0.196	Compton (C2)

Although not comparing the results with experiment, an attempt is made to present the argument in such a manner that a comparison with data is readily possible. Indeed, a comparison has been made with recent data on the F center in KCl. The agreement on the whole is satisfactory. It is hoped that in the next few years sufficient experimental data will become available to justify and amplify the theory presented here.

The first theory of the shape of an absorption band in a solid is due to Smakula (S2), who assumed that a trapped electron behaves like a damped oscillator in a dielectric medium. Nothing in this theory explains the temperature dependence of the damping term. The first attempt to associate the width theoretically with the thermal vibration was done independently by Muto (M13), by Huang and Rhys (H5), and by Pekar (P1) (also P2, 3, and 4). Huang-Rhys and Pekar used Fröhlich's model and assumed that the electron interacts only an infinitesimally small amount with all the longitudinal optical modes. Thus they limited themselves to first order terms in the various expansions, as did Muto.

Pekar obtained two expressions for the shape; it should be nonsymmetric at low temperatures but Gaussian at high temperatures. He employed the reststrahlen frequency combined with the Lyddane-Sachs-Teller (LST) relation to determine the longitudinal frequency (B2, p. 82–128). He also made a comparison of his theory with Mollwo's data (M10).

O'Rourke (O1) made a fundamental extension of the previous calculations, eliminating the restriction that

† Half-width is defined in Appendix I.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

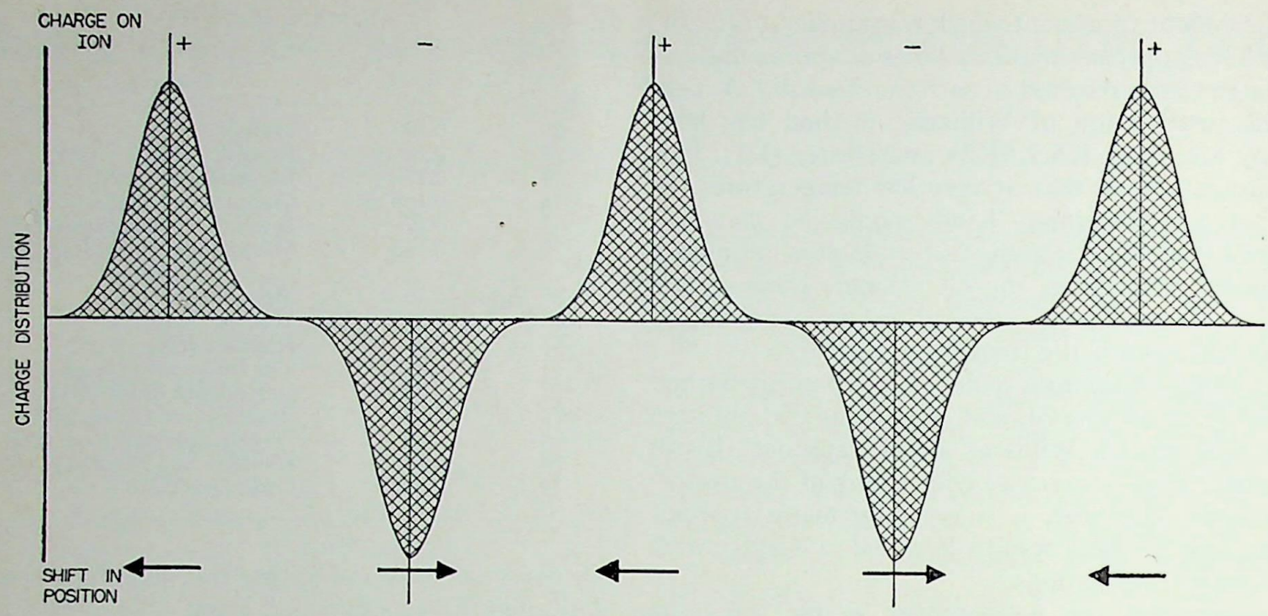


FIG. 1. Schematic diagram showing the effects of trapping an electron at an imperfection.

the electron-phonon interaction must be infinitely small. His most general expressions are limited by the assumption that electronic transitions do not cause frequency changes. To use his expressions one must assume that a single "effective" vibrational frequency exists. His results apply to the Fröhlich Hamiltonian, the CC model, as well as to an LM model, if a single effective frequency exists. Pekar's expansions may therefore be combined with the CC or LM model to calculate the shape rigorously at very low temperatures. We do this and show that the shape is not the one predicted on the basis of the Williams' approximation. Present low-temperature data indicate that the more rigorous treatment is correct.

At about the same time, Lax (L2) developed some completely general relations between the moment of the distribution and the temperature. Meyer (M8 and 9) extended the development of Lax, giving a general approach to the problem. This method is not limited to Fröhlich's Hamiltonian which was employed by Meyer. For the calculation to be valid, it is necessary to use an orthogonal set of eigenfunctions. We show here that this is true. Further, we indicate that Meyer omits some important terms.

Meyer also made a serious attempt to analyze Mollwo's data by means of the new theories. His results are most impressive and by themselves give a good proof of the reliability of his calculation and a validity of the LST relation. This is not confirmed by more recent measurements (see Table I). The method of analysis to be employed is of fundamental importance. An alternate analysis of Mollwo's data indicates that the effective frequency is 3.2×10^{12} sec^{-1} for KCl. The LST relation gives 6.3×10^{12} sec^{-1}. This difference suggests that Fröhlich's model does not apply to the F center in KCl. Other analyses (R1, K4) support this view. The means of analysis is the important difference between Meyer's conclusion and those obtained recently.

This field has been reviewed by Pekar (P3, P4), Lax (L2), Meyer (M9), Klick and Schulman (K2a), and Dexter (D1). Pekar's reviews, though out of date, are extremely important contributions to the field. Lax's review concerns itself primarily, though not entirely, with Williams' approximation. Meyer's review is limited to Fröhlich's model and omits some important developments made by Pekar. Klick and Schulman consider only Williams' approach. Dexter considered both aspects of the problem, applying the quantum mechanical treatment to Fröhlich's model and Williams' approximation to the CC model.

This paper concentrates on the pure quantum mechanical approach and treats all models. Some aspects of the problem are considered in greater detail than has been done previously. These aspects are most important since they can be checked experimentally.

The theory is not developed in full; instead, results are quoted from the recent book of Born and Huang (B2), which is cited by the letter B. Those parts of the theory which have not been presented in book or review article form are more fully discussed. Although the theory applies to any imperfection with two or more bound states, the development is closely related to the F center, since this has been studied in greatest detail.

Part I. The Hamiltonian

2. PHYSICAL CONCEPTS INVOLVED IN ELECTRON-PHONON INTERACTIONS

The trapping of an electron at a site has three effects. These can be intuitively understood by means of Fig. 1, which is a schematic representation of the nuclei

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

(shielded by the inner electrons) for a one-dimensional polar crystal.

The nuclei of the solid are assumed to be initially distributed uniformly. When an electron is trapped nearby (to the left of the diagram) the following effects occur.

(1) The electron induces a force on the nuclei, causing a nonuniform shift of the rest positions, as indicated by the arrows. This action is very much like a compression (or expansion) of a spring and leads to the storing of energy.

(2) The forces between the nuclei are slightly modified, leading to a shift in the angular frequencies of the lattice, $\Delta\omega$. At absolute zero the probable position of an ion is given by a Gaussian curve if an Einsteinian model is used (no coupling between ions). The trapping causes a broadening or narrowing of the distribution. The Einsteinian solid is used here only to illustrate the effect.

(3) Finally, there is a third effect similar to a photon-electron interaction—the electron's self-energy. The electrons and nuclei readjust their probable distributions so that the total energy is at a minimum. This effect may be visualized by means of time dependent perturbation theory leading to electron-phonon transitions, or by time independent perturbation theory (or even variational methods) leading to wave functions which are no longer single products of electronic and nuclear wave functions. This is referred to as the electron-nuclear correlation.

These effects also occur when an electron redistributes itself, i.e., acquires a new wave function by jumping from one bound state to another, or by transitions to the conduction band. Item (3) is similar to phenomena which occur in electrodynamics, while (1) and (2) arise in systems made of light and heavy particles. The relative importance of these effects depends on the forces within the solid. In many solids, effects (1) and (2) are so small that phonon-electron interaction problems resemble problems in electrodynamics. This is true in metals and in some semiconductors. In general, the phonon-electron problem has a different character because of effects (1) and (2). Our interest here is the difference between phonon-electron and photon-electron interactions; hence, effect (3) is not considered in detail. Usually the effects are considered independently.

Treatments of the correlation problem appear several places and take various forms. Born and Huang (B, p. 166) considered it briefly in a general manner, while Haken (H1) studied its effect in a special one-dimensional model.

Simpson (S3) and Pekar (P3) considered effect (1) in their calculations on the F center. Williams (W2) in an ambitious and detailed study calculated the displacements which occur when a Tl^+ ion replaces a K^+ ion in a KCl lattice. Williams finds a change in the second derivative of his CC plots which is related to $\Delta\omega$. Klick and Schulman (K2a) have attempted to compute $\Delta\omega$ from experimental data.

The idea of a large frequency shift was invoked by Seitz (S1) to explain the observations of Dutton and Maurer (D4). They measured a thermal activation energy for the V_1 band in KBr of 0.23 ev compared to an optical activation energy of 3.0 ev. These ideas have some theoretical difficulties. A similar problem exists for the F' center whose optical activation energy in KBr is 1.3 ev compared to a thermal activation energy of 0.29 ev (D4).

A problem which involves all three effects is an electron at the bottom of the conduction band in a polar solid. The electron interacts strongly with the surrounding ions. Due to the large difference in the velocity of the electrons and the nuclei, the forces on the nuclei depend on the average distribution of the electron, thus the electron displaces the surrounding ions. This causes effects (1) and (2) and results in a potential well. The well in turn traps the electron. The electron will also be influenced by the polar modes of vibration about the well. This is similar to the interaction between an electron and the zero-point electromagnetic vibrations. The presence of the well may be referred to as self-trapping,‡ while the electron-phonon interaction may be called the polaron effect. In reality, they cannot be separated, although at present the connection between the two is not completely clear.

3. FORMAL THEORY

The complete Hamiltonian for a solid of unit volume is

$$H=T_e+T+V(\mathbf{R}). \tag{3.1}$$

Here T_e is the kinetic energy operator for all the electrons, T is the kinetic energy operator for the N nuclei, and V is the total potential energy of the system. V depends on the position of all the electrons (denoted by $\mathbf{r}$ when explicitly stated) and all the nuclei (denoted by $\mathbf{R}$). We include in V implicitly any point imperfection such as a missing ion. The notation of (B) will be used, namely, the coordinate of a nucleus is $X_\alpha(k)$ ($\alpha=1, 2, 3$ and $k=1, 2\cdots N$). The explicit expression for T is

$$T=-(\hbar^2/2)\sum_k (1/M_k)\nabla_k^2, \tag{3.2}$$

where all the nuclei need not have the same mass, M_k.

The Born-Oppenheimer technique assumes that for a fixed value of $\mathbf{R}$ the eigenfunctions, $\varphi_n(\mathbf{R})$, and eigenvalues, $\epsilon_n(\mathbf{R})$, associated with the operator

$$h_e(\mathbf{R})=T_e+V(\mathbf{R}) \tag{3.3}$$

are known.

‡ The assumption has been made that the self-trapping concept implies a stationary distortion of the lattice. There is no reason for this assumption. Indeed, Seitz and the author (M6) suggested at an early date that a self-trapped electron might move through the lattice. The work of Castner and Kanzig (C1) implies that a self-trapped hole has relatively little mobility.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

Next, one uses the Hamiltonian

$$h_v(\mathbf{R}_n)=T+\epsilon_n(\mathbf{R})-\epsilon_n(\mathbf{R}_n) \tag{3.4}$$

to obtain a second set of eigenfunctions χ and eigenvalues ϵ_v. The point $\mathbf{R}_n$ of (3.4) is defined by the relations

$$[\partial\epsilon_n/\partial X_\alpha(k)]_{\mathbf{R}_n}=0. \tag{3.4a}$$

To terms in $[\Delta X_\alpha(k)]^3$ we may write (3.4) in the form

$$h_v=T+\tfrac{1}{2}\sum_{\alpha,\beta,k,l}\left[\frac{\partial^2\epsilon_n}{\partial X_\alpha(k)\partial X_\beta(l)}\right]_{\mathbf{R}_n}\Delta X_\alpha(k)\Delta X_\beta(l). \tag{3.5}$$

Here $\Delta X_\alpha(k)$ is the displacement relative to $\mathbf{R}_n$. One may introduce normal coordinates into (3.5) (B, p. 173) which are linear functions of the $\Delta X_\alpha(k)$. The result is

$$h_v=\tfrac{1}{2}\sum_j(p_j^2+\omega_j^2q_j^2), \tag{3.6}$$

where

$$p_j=(\hbar/i)(\partial/\partial q_j), \tag{3.6a}$$

and j goes from 1 to 3 N. The q's have the dimension of length times (mass)$^{\frac{1}{2}}$ and the ω's are given by the secular determinant,

$$\left|\frac{1}{(M_kM_l)^{\frac{1}{2}}}\left[\frac{\partial^2\epsilon_n}{\partial X_\alpha(k)\partial X_\beta(l)}\right]_{\mathbf{R}_n}-\omega^2\delta_{\alpha\beta}\delta_{kl}\right|=0. \tag{3.7}$$

Here δ is the Kronecker delta.

The individual term in (3.6) corresponds to a simple harmonic oscillator with an eigenfunction, χ_{nj}, and an eigenvalue, $\hbar\omega_j(v_j+\frac{1}{2})$. The eigenfunctions associated with h_v are products of such functions and are denoted by χ_n. n corresponds to an electronic state and appears through the use of $\epsilon_n(\mathbf{R})$, and v_j is the vibrational quantum number of the jth mode. v_t is the sum of the v_j's. Since we are mainly concerned with two states, n will at times be replaced by g (ground) and u (upper). Usually the subscript n will not be required. φ and χ will be used for the ground state while φ' and χ' will be employed for the excited state.

To complete the notation, the following definitions are made:

$$\langle\varphi_{n'}|\mathbf{O}|\varphi_n\rangle=\int\varphi_{n'}{}^*\mathbf{O}\varphi_n d\tau, \tag{3.8}$$

where the integration is over all positions of the electrons, and

$$\{\chi_{n'}|\mathbf{O}|\chi_n\}=\int\chi_{n'v'}{}^*\mathbf{O}\chi_{nv}d\mathbf{R}, \tag{3.9}$$

where the integration is over all positions of the nuclei. $\mathbf{O}$ is any operator. It will be omitted when $\mathbf{O}=1$.

The essential problem is to find approximate eigenfunctions for the total Hamiltonian (3.1). By using the foregoing equations, we may define two approximations:

(1) the static—$\Psi(S)=\varphi(\mathbf{R}_n)\chi$ (3.10)

or

(2) the adiabatic—$\Psi(A)=\varphi(\mathbf{R})\chi$. (3.11)

Neither (3.10) nor (3.11) is a complete solution of (3.1). For some problems one desires to split H into a zeroth-order Hamiltonian and a residual which represents a small perturbation. The way this is done depends on the form of Ψ employed.

For the static approximation

$$H_0(S)=h_e(\mathbf{R}_n)+h_v(\mathbf{R}_n) \tag{3.12a}$$

$$H_1(S)=V(\mathbf{R})-V(\mathbf{R}_n)$$

$$-\tfrac{1}{2}\sum_{\alpha,\beta,k,l}\left[\frac{\partial^2\epsilon_n}{\partial X_\alpha(k)\partial X_\beta(l)}\right]_{\mathbf{R}_n}\Delta X_\alpha(k)\Delta X_\beta(l). \tag{3.12b}$$

On the other hand, in the adiabatic approximation,

$$H_0(A)\varphi(\mathbf{R})\chi=\chi h_e(\mathbf{R})\varphi(\mathbf{R})+\varphi(\mathbf{R})T\chi \tag{3.13a}$$

$$H_1(A)\varphi(\mathbf{R})\chi$$

$$=-\frac{\hbar^2}{2}\sum_k\left[\frac{2}{M_k}\nabla_k\varphi\cdot\nabla_k\chi+\frac{1}{M_k}\chi\nabla_k{}^2\varphi(\mathbf{R})\right]. \tag{3.13b}$$

$H_0(A)$ assumes that $\varphi(\mathbf{R})$ permutes with T.

An elegant method of arriving at (3.12) was given many years ago by Born and Oppenheimer (B3). This treatment is reproduced in a simpler form by (B, pp. 166–173), who have also found a systematic way to obtain the adiabatic approximation. In the latter case, the χ's contain higher-order anharmonic terms. The Born-Oppenheimer technique involves several steps. Equation (3.4a) is a particularly important one, since it is involved in h_v. Limitations on this technique are discussed in Sec. 5(b).

To obtain the eigenvalues of (3.13a) we may proceed as follows:

$$\begin{aligned}H_0(A)\varphi(\mathbf{R})\chi&=\varphi(\mathbf{R})[\epsilon_n(\mathbf{R})+T]\chi\\&=\varphi(\mathbf{R})[\epsilon_n(\mathbf{R}_n)+h_v]\chi\\&=[\epsilon_n(\mathbf{R}_n)+\epsilon_v]\varphi(\mathbf{R})\chi,\end{aligned} \tag{3.14a}$$

where (3.4a) and (3.5) have been used. The eigenvalues $H_0(S)$ and $H_0(A)$ are equal since

$$H_0(S)\varphi(\mathbf{R}_n)\chi=[\epsilon_n(\mathbf{R}_n)+\epsilon_v]\varphi(\mathbf{R}_n)\chi. \tag{3.14b}$$

In view of (3.14a) and (3.14b) we define the total energy of the system as $E_{nv}=\epsilon_n(\mathbf{R}_n)+\epsilon_v$. We shall again omit the subscripts, using E and E'.

The method of splitting H into parts is based on a physical intuition and the fact that the nuclear mass is much heavier than the electron mass. Mathematically, the splitting has an unfortunate aspect, since some of the nicest mathematical properties do not hold for the H_0's.

One may show (M4) that (1) the static eigenfunctions do not form an orthogonal complete set while (2) the adiabatic eigenfunctions form an orthogonal set with respect to $\mathbf{r}$ and $\mathbf{R}$ space.

The latter can be shown by noting that the $\varphi(\mathbf{R})$'s form a complete set of functions in $\mathbf{r}$ space for a fixed $\mathbf{R}$,

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

and that the χ's form a complete set in $\mathbf{R}$ space for a fixed n. Further, we note

$$\{\langle\varphi_{n'}(\mathbf{R})\chi_{n'v'}|\,\varphi_n(\mathbf{R})\chi_{nv}\rangle\}=\{\chi_{n'v'}\langle\varphi_{n'}(\mathbf{R})|\,\varphi_n(\mathbf{R})\rangle\chi_{nv}\}$$
$$=\{\chi_{n'v'}|\,\delta_{nn'}|\chi_{nv}\}=\delta_{nn'}\delta_{vv'}. \quad (3.15)$$

(3) It can be shown that (B, p. 189)

$$\left[\frac{\partial\epsilon_n}{\partial X_\alpha(k)}\right]_{\mathbf{R}_n}=\left\langle\varphi_n(\mathbf{R}_n)\left|\left[\frac{\partial V}{\partial X_\alpha(k)}\right]_{\mathbf{R}_n}\right|\varphi_n(\mathbf{R}_n)\right\rangle \quad (3.16)$$

(Feynman's Theorem).

For a perfect ionic crystal the lowest eigenvalue of h_e is the total interionic potential, i.e., the coulombic, repulsive, and high-order terms, from which one determines the normal modes by well-developed methods. In the case of interest, all cells in our solid are not equivalent, since there is a point imperfection. In the case of the F center there is a missing negative ion with a trapped electron. Mazur, Montroll, and Potts (M2) have shown that the modes for such a crystal are not identical to those of a perfect lattice. A missing ion creates local modes surrounding the imperfection. These are the vibrations of greatest interest to our problem. In a general way the motion of the ion right around an imperfection is determined primarily by the local modes (Bjork, B1). The object here is not to describe these modes in detail but to indicate what effects electron transitions have on them. The exact relation between local and crystal modes is an extremely complex problem. One of the best ways to study the local modes is to measure their frequencies indirectly. This can be done from measurements of the thermal broadening of absorption and emission bands (Part II).

(a) Ground State

The development so far is completely general. We now use this development for a crystal with an electron trap. Assume that there are two (or more) bound states. Every state generates its own ϵ_n, ω_j's, and q_j's. We start the calculation by using the values appropriate for the ground state. To stress this point we write ϵ_g for ϵ_n; $\omega_j(g)$ for ω_j; $h_v(g)$ for h_v, etc. We shall retain the notation q_j and p_j, however.

(b) Upper State

We could make all the equations apply to the upper state by simply replacing the subscript g by u. The p's and q's of (3.6), however, will be different. The object is to find relations between the two sets; hence, we proceed in a different manner. In this case, the eigenvalue of h_e is $\epsilon_u(\mathbf{R})$ and the trapping energy may be defined as

$$\Delta\epsilon(\mathbf{R})=\epsilon_u(\mathbf{R})-\epsilon_g(\mathbf{R}). \quad (3.17)$$

$\Delta\epsilon$ is approximately equal to the energy the trapped electron acquires when it is excited from one state to another (the nuclear coordinates remain at $\mathbf{R}$). Actually, the difference in energy is related to all the electrons rather than just one.

Since we are interested in the q's, $\epsilon_u(\mathbf{R})$ is expressed as

$$\epsilon_u(\mathbf{R})=\epsilon_g(\mathbf{R})+\Delta\epsilon(\mathbf{R})=\epsilon_g(\mathbf{R}_g)+\tfrac{1}{2}\textstyle\sum_j\omega_j^2(g)q_j^2$$
$$+\Delta\epsilon(\mathbf{R}_g)+\textstyle\sum_j\epsilon_jq_j+\tfrac{1}{2}\sum_{jk}\epsilon_{jk}q_jq_k+\cdots, \quad (3.18)$$

where

$$\epsilon_j=[\partial\Delta\epsilon/\partial q_j]_{\mathbf{R}_g} \quad (3.18a)$$

and

$$\epsilon_{jk}=[\partial^2\Delta\epsilon/\partial q_j\partial q_k]_{\mathbf{R}_g}. \quad (3.18b)$$

The dimension of ϵ_j is (mass)$^{\frac{1}{2}}$ (length) (time)$^{-2}$ while the dimension of ϵ_{jk} is simply (time)$^{-2}$. The second term in (3.18) arises because $\epsilon_g(\mathbf{R})-\epsilon_g(\mathbf{R}_g)=\frac{1}{2}\sum_j\omega_j^2(g)q_j^2$, which follows from the definition of the q's.

At this point the theory runs into some difficulties. Equation (3.4) assumes that terms in $\Delta X_\alpha(j)\Delta X_\beta(k)\times\Delta X_\gamma(l)$ are so small that they can be neglected. If the q's are determined by thermal vibrations, they will be small at low temperatures. Equation (3.18) includes, however, the displacement which occurs during a transition, i.e., the difference between $\mathbf{R}_u$ and $\mathbf{R}_g$ defined by

$$[\partial\epsilon_g/\partial R]_{\mathbf{R}_g}=0 \quad (3.19a)$$

and

$$[\partial\epsilon_u/\partial R]_{\mathbf{R}_u}=0. \quad (3.19b)$$

Actual calculations of the difference between the $\mathbf{R}$'s have been made on some simple models. $\mathbf{R}_g-\mathbf{R}_u$ is about 0.3 A for the nearest neighbor in the case of a Tl^+ trap in KCl (W2) and 0.06 A for the self-trapped electron (M6). These results are to be compared to the root-mean-square displacement of an atom (Einsteinian model with a mass equal to 30 times that of the proton and frequency of 5×10^{12}). For $n=0$ (the quantum vibrational number), it is 0.06 A and increases to 0.13 A for $n=2$. X-ray measurements of the root-mean-square amplitude of vibration in NaCl at 86°K gives 0.15 A for Na^+ and 0.13 A for Cl^- (L3, p. 50). These large values must be attributed to the lower-frequency modes. The self-trapping calculations showed that the first anharmonic term was of no importance, due to a fortuitous geometry. However, this is not a general conclusion. If we include additional terms in (3.18), there seems to be no apparent simple way to relate the normal modes in the upper and lower states. Therefore, we consider (3.18) to be exact. For shallow traps and for any trap in nonionic crystals, this assumption must be *good*. It must also be a *fair* approximation for deeper traps in ionic crystals, such as the F center.

It is further assumed that the modes are nondegenerate. To those familiar with Fröhlich's phenomenological treatment, this may seem like a radical assumption indeed. One must remember that Fröhlich's approximation is not supported by detailed calculations

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

(M3). Secondly, we may be dealing to a large extent with local modes. This is supported by the analyses on the shape of the absorption bands. These are in the forbidden gap and cluster about a single value. There is no reason, however, to assume that the local modes are actually degenerate, and the above assumption is a natural one at the present stage of development. This assumption probably does not affect any major conclusions of this paper.

Now the transformation

$$q_j = q_j' + \sum_k{}' \frac{\epsilon_{jk}}{\omega_k^2(g) - \omega_j^2(g)} q_k' \tag{3.20}$$

is made. The prime on the sum indicates that $j \neq k$. Neglecting terms of order $\epsilon_{ij}\epsilon_{kl}$ or $\epsilon_j\epsilon_{kl}$ for $(k \neq l)$, (3.20) gives an orthogonal transformation. To the same approximation we write

$$\epsilon_u(\mathbf{R}) = \epsilon_g(\mathbf{R}_g) + \Delta\epsilon(\mathbf{R}_g) + \tfrac{1}{2}\sum_j \omega_j^2(u)(q_j')^2 + \sum_j \epsilon_j q_j' \tag{3.21}$$

$$\omega_j^2(u) = \omega_j^2(g) + \epsilon_{jj}. \tag{3.21a}$$

The new equilibrium position is obtained from (3.19b). Since the ΔX's are linear functions of the q's, (3.19b) gives

$$(\partial/\partial q_j')\epsilon_u = \omega_j^2(u) q_j' + \epsilon_j = 0. \tag{3.22}$$

The shift in the equilibrium position is

$$q_j' = -\epsilon_j/\omega_j^2(u). \tag{3.23}$$

Thus, the final transformation

$$Q_j = q_j' + \epsilon_j/\omega_j^2(u) = q_j' - \Delta q_j \tag{3.24}$$

gives the energy of the ground state in the form

$$\epsilon_u(\mathbf{R}) = \epsilon_g(\mathbf{R}_g) + \Delta\epsilon(\mathbf{R}_g) - \tfrac{1}{2}\sum_j \epsilon_j^2/\omega_j^2(u) + \tfrac{1}{2}\sum_j \omega_j^2(u) Q_j^2 \tag{3.25}$$

$$= \epsilon_u(\mathbf{R}_u) + \tfrac{1}{2}\sum_j \omega_j^2(u) Q_j^2. \tag{3.25a}$$

By direct substitution one may show that

$$\sum_j \left(\frac{\partial}{\partial q_j}\right)^2 = \sum_j \left(\frac{\partial}{\partial q_j'}\right)^2 = \sum_j \left(\frac{\partial}{\partial Q_j}\right)^2, \tag{3.26}$$

hence, the transformation does not modify T.

The first three terms on the right of (3.25) appear in the eigenvalue of h_e. The derivation serves to show the relation between the q's and Q's. This is of essential importance and not the final form of ϵ_u.

This concludes the formal treatment. The assumptions which have been made are now summarized to emphasize their limitations. These are that

(3.27)

(1) all cubic terms can be neglected;

(2) terms in $\epsilon_{ij}\epsilon_{kl}$ and $\epsilon_j\epsilon_{kl}$ for $k \neq l$ can be neglected;

(3) the normal modes are nondegenerate.

In view of the work of Montroll and co-workers, we know that lattice perturbation can localize some of the modes. This effect does not appear above. The transformation (3.20) without the neglect of higher-order terms (item 2) may cause a radical change in the modes. This change most probably does not occur however during an electronic transition.

The third assumption probably is not a fundamental one, although the first two may be. The problem of degenerate modes is that the q's and p's of Eq. (3.6) are not unique. Consider two degenerate modes q_1 and q_2. The substitution $q_1'' = q_1 + q_2$ and $q_2'' = q_1 - q_2$ will not change the required form of the Hamiltonian, and the q'' modes are equally satisfactory. The additional terms which occur in the upper state put a requirement on the possible linear combination of the q's. These can be obtained from argument similar to those used in degenerate quantum mechanical perturbation theory. If we use the proper definition of the q's, the degenerate problem is similar to the nondegenerate one. For this reason, the author feels that item 3 is not an essential restriction. The author has not carried through the details of this argument. Certainly the transformations (3.20) and (3.24) allow one to make a simple synthesis of many theoretical ideas at a time when it is most needed from the experimental point of view. Hence, we assume their validity without any further ado.

For a point imperfection we may expect two types of ϵ_j. The ϵ_j's for the local modes will be especially large in solids with some polar characteristics. This leads to the LM approximation. The ϵ_j's for individual longitudinal modes are very small; the net effect, however, may be large. This leads to Hamiltonians similar to those employed by Fröhlich. For these modes the ϵ_{jj}'s must be extremely small and the transformation (3.20) need not be used.

Item 1 of Sec. 2 enters into the problem through (3.19b) or (3.24), item 2 through $\omega_j(g)$ or (3.21a), while item 3 arises through (3.12b) or (3.13b) depending on the form of Ψ employed.

The technique used on the F center by Huang and Rhys and by Pekar assumes that ϵ_j is of the order of $1/N$. This does not hold for local modes; hence the expansion techniques are questionable. O'Rourke does not require that $\epsilon_j \to 0$, in the first part of his paper, thus includes localized modes.

The division made above regarding the types of modes is not completely satisfactory since the Fröhlich description is somewhat oversimplified. The usual classification of modes, i.e., acoustical *versus* optical, transverse *versus* longitudinal, apply only to the long modes in polar crystals (M3). The majority of the modes actually have a wavelength of the order of several interionic distances and do not fit into the above classification. The experimentalist should remember that the LST relation is not completely reliable, and, therefore, should not take it too seriously. The theorist should attempt to formulate his equations in such a

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

manner that he does not rely on the extreme simplification involved in this relation. As our understanding of the problems treated here increases (this involves further experimental data), we shall undoubtedly obtain much more reliable information regarding the vibrational modes in solids.

4. PHYSICAL INTERPRETATION OF OPTICAL AND THERMAL TRANSFORMATIONS

A physical interpretation of Sec. 3 is now attempted. It starts with a one-dimensional Franck-Condon diagram which will subsequently be elaborated. Next, the various activation energies are defined and their interrelations discussed. The results are applied to the V_1 center.

We start by considering the following simple model: The temperature is 0°K. Only one mode is affected by the transition. This implies that only one ϵ_j is different from zero. In Fig. 2 we plot ϵ_u and ϵ_g as a function of a coordinate q_j. It is convenient, for now, to set $\epsilon_{jj}=0$; hence $\omega_j(u)=\omega_j(g)$. Although ϵ is an eigenfunction of h_e, it is useful to divide it into two parts, namely, $\epsilon_g(\mathbf{R}_0)=\epsilon_g(0)$, which is considered as an electronic energy of the ground state, and $\epsilon_g(q_j)-\epsilon_g(0)$, which is considered as a lattice potential energy. This division is justified by Eqs. (3.14a and b). According to the simplified Franck-Condon principle, the system may absorb a photon and jump from point a to point b on ϵ_u. The use of the Born-Oppenheimer technique assumes that the electron has sufficient kinetic energy to occupy all the space required by $|\varphi'|^2$ before there has been an appreciable displacement of q_j from 0. ϵ_g has been replaced by ϵ_u, giving an additional "force" $-\epsilon_j$ on q_j. In the case considered in Fig. 2 $\epsilon_j=-\omega_j^2\Delta q_j$. Hence the total potential energy of the lattice relative to point b is $\frac{1}{2}\omega_j^2 q_j^2+\epsilon_j q_j$ and its minimum appears at $q_j=-\epsilon_j/\omega_j^2$ in agreement with (3.23). As soon as the force on the q_j'th mode has changed, its equilibrium value has as well, although the ions are not at $-\epsilon_j/\omega_j^2$.

During the shift from b to c, $\Delta\epsilon$ *decreases* by $-\epsilon_j\Delta q=(\epsilon_j^2/\omega_j^2)$, while the potential energy of the lattice *increases* by $\frac{1}{2}\epsilon_j^2/\omega_j^2$, thus giving a total decrease of $\frac{1}{2}\epsilon_j^2/\omega_j^2$. Hence the third term of (3.25)* is composed of two effects, the decrease of electronic energy and the increase of lattice energy. The simplicity of the interpretation arises because of the assumption that the restoring forces are not affected by the jump of the electron in the trap.

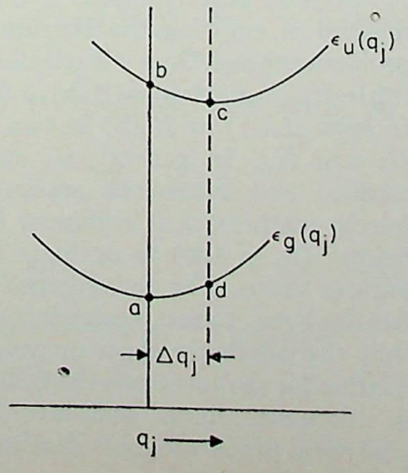


FIG. 2. Plots of ϵ_g and ϵ_u against q_j.

Quantum mechanics (still assuming that $\epsilon_{jj}=0$) only slightly modifies the picture. The stationary states are represented by horizontal lines since normal vibrations have kinetic as well as potential energy. The most probable transitions arise when the classical turning points of the upper and lower states occur for the same value of q_j. While this is the most probable, it is not the only one; indeed, the probability of "nonvertical" transitions explains the shape of the emission and absorption bands in solids.

The theory developed in Sec. 3 is now used to define the various types of activation energies. Neglecting the broadening due to the emission or absorption of phonons, the zero-point energy, and limiting our consideration to 0°K, we may define the following: activation energy for emission of a photon—E^e—$(c-d)$; activation energy for absorption of a photon—E^a—$(a-b)$; thermal activation energy which is the same for emission and absorption—E^t—$(a-c)$. The a, b, c, and d refer to Fig. 2.

As has been stressed previously, the Stokes' shift, which is the difference between E^e and E^a, is related to the third term of (3.25). After an absorption (from a to b), the force $-\epsilon_j$ is added, and the system relaxes by going to c. For $\epsilon_{jj}=0$, the stored energy after an emission equals the stored energy after an absorption, i.e.,

$$\epsilon_g(-\epsilon_j/\omega_j^2)-\epsilon_g(0)=\epsilon_u(0)-\epsilon_u(-\epsilon_j/\omega_j^2). \qquad (4.1)$$

In (4.1) the arguments of ϵ_u refer to values of q not Q. Summarizing, the absorption activation energy includes the readjustment energies in both the upper and ground states; the emission activation energy excludes both, while the thermal includes the readjustment energy in only one state.

Although (4.1) was derived by considering the displacement of only one mode, it is quite general because ϵ_g and ϵ_u are functions only of the q's and not of the path taken. The displacements can be compared one by one to show that for an actual trap the total stored energy is the same in both states. In general then, one may write

$$E^e+2\delta E=E^t+\delta E=E^a \qquad (4.2)$$

where δE is the readjustment energy $(\frac{1}{2})\sum_j \epsilon_j^2/\omega_j^2$, or

$$E^t=E^a-(1/2)(E^a-E^e)=(1/2)(E^a+E^e). \qquad (4.2a)$$

If $E^e=0$, $E^t=\frac{1}{2}E^a$; hence, in general

$$E^t\geqslant(1/2)E^a \quad (\text{for } \epsilon_{jj}=0). \qquad (4.2b)$$

In some problems we may be able to introduce an

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

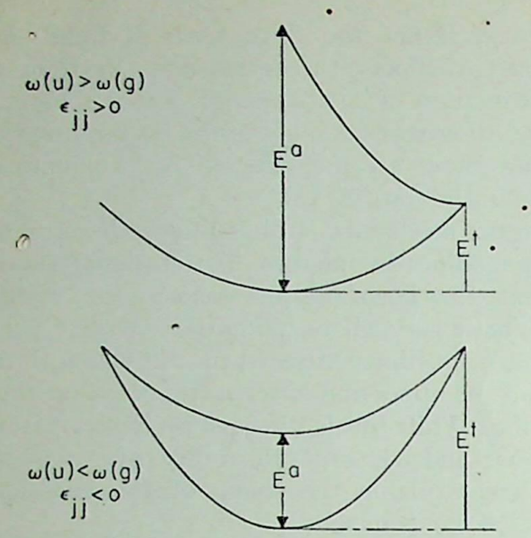


FIG. 3. Effects of ϵ_{jj} on E^a and E^t.

effective frequency (Secs. 9 and 10). In this case

$$\delta E=\tfrac{1}{2}\sum_j \epsilon_j^2/\omega_j^2=\hbar\omega\{\tfrac{1}{2}\sum_j \epsilon_j^2/\hbar\omega_j^3\}=\hbar\omega S \qquad (4.3)$$

where ω is the effective frequency, defined by (4.3), and

$$S=\tfrac{1}{2}\sum_j \epsilon_j^2/\hbar\omega_j^3, \qquad (4.3a)$$

the Huang-Rys factor. One may measure S directly for some transitions (Sec. 11). It is the ratio of the stored energy at b to the average effective phonon energy. For the F center in KCl, $S\approx 30$ (K4).

A relaxation of (4.2) occurs when $\epsilon_{jj}\neq 0$. When $\epsilon_{jk}(j\neq k)$ is small compared to ϵ_{jj}, the effect of the second derivative is to shift the frequencies of the modes. One may speak of a frequency shift superimposed on a displacement of the equilibrium position. Now (4.2b) does not apply, and E^t can be much smaller than $\frac{1}{2}E^a$. In other cases it can be even larger than E^a, as illustrated in Fig. 3,§ where only one mode is considered. If $E^t\ll E^a$, ϵ_{jj} must be of the order of ω_j^2. For $\epsilon_{jj}<0$, one obtains the strange case where $E^a<E^t$. However, the author knows of no such case in nature.

In the self-trapping calculations made by Seitz and the author (M6) for NaCl, items 2 and 3 of Sec. 2 were not considered; hence, $E^t=\frac{1}{2}E^a$. This gives 0.34 ev for E^t, not the value originally reported. The error arises from an erroneous formulation of Feynman's theorem (Eq. 3.16). The agreement between the calculation of Fröhlich, Pelzer, and Zienau (F3) and the reported value of E^t (i.e., 0.13 ev) must be fortuitous. The author believes that higher-order terms than ϵ_j must be included. Further, the mobility of the trap has to be included; hence the value 0.34 ev is a very preliminary one.

In view of the values of E^t and E^a for the F' center, the ϵ_{jj}'s must almost equal ω^2's for the local modes which cluster about negative-ion vacancies. The upper F' state corresponds to an F center and an electron in the conduction band. The lower state corresponds to two electrons in a negative-ion vacancy. To the approximation of interest here we may assume that the free electron does not interact with the modes. $\Delta\epsilon(\mathbf{R})$ of Eq. (3.17) is simply the binding energy of the second electron in the F' center. Since there is a radical difference in the upper and lower states, the assumption of $\epsilon_{jj}\approx\omega_j^2(g)$ seems plausible. The author finds it hard to believe that a corresponding situation exists for a hole trapped at a positive-ion vacancy. One assumes that it possesses two bound states. This is Seitz's model for the V_1 band in KCl and KBr. The hole must be spread out similar to an electron in the F center. Transitions of the hole from an excited to a ground state could influence the vibrational frequencies but hardly enough to change the activation energy from 3.0 ev to 0.23 ev. Our present knowledge of the V centers is so meager that the possibility of an alternate interpretation of the experiment should not be overlooked.

§ *In Fig. 3, the problem of the intersection of the potential curves has not been considered.*

5. PROPERTIES OF THE TOTAL HAMILTONIAN

In Sec. 3 a theory was developed which is completely general, provided one may use the Born-Oppenheimer technique. In this section we compare this development to alternate formulations of the problem in order to contrast the ideas. We also compare the development of Sec. 3 with some problems in electrodynamics, in particular, to contrast the behavior of phonons and photons.

In complex systems one likes to split the Hamiltonian into parts which interact only slightly. Thus, in electrodynamics, it is divided into the radiation field $H_{\rm rad}$, the part related to the charged particles $H_{\rm el}$, and the interaction term $H_{\rm int}$, which relates the electrons to the photons. One usually develops expressions for $H_{\rm rad}$ and $H_{\rm el}$ without regard to each other. The introduction of an electron in a vacuum does not affect $H_{\rm rad}$, nor does the introduction of black-body radiation affect $H_{\rm el}$. When one considers a trapped electron in an electromagnetic field one adds $H_{\rm int}$ to $H_{\rm rad}+H_{\rm el}$ without modifying either.

The solid state problem is more complex, and the method of splitting the Hamiltonian depends on the problem at hand. Once Ψ is selected, the term corresponding to $H_{\rm int}$ is given by (3.12b) or (3.13b). The problem is to split $H_0(A)$ or $H_0(S)$ in two. These we denote by H_e and H_L. In general, H_e and H_L are intimately coupled and cannot be separated. One may use a purely mathematical argument like Born and Oppenheimer (B, p. 166) to evaluate the importance of the various terms. The convergence of the approximation has not been studied, however. Crude calculations show that the adiabatic approximation must be used with caution for shallow traps (M4). The splitting procedure is discussed in a semi-intuitive manner; further analytical proofs are most desirable.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

(a) Trapped Electron

For an electron trapped at a point imperfection, we may use (3.1) or a form which is related more closely to the way actual calculations are made. For the latter purpose we define the impurity center Hamiltonian. With proper care one may make the treatment completely rigorous. Thus we write

$$H=t+T+V_L(\mathbf{R})+V_I(\mathbf{R})+V_e(\mathbf{R},\mathbf{r}). \quad (5.1)$$

Here t is the kinetic energy of the trapped electron; V_L+V_I is the interionic or interatomic potential; $V_e(\mathbf{R},\mathbf{r})$ is the potential energy due to an excess electron at $\mathbf{r}$ when the ions (or atoms) are at $\mathbf{R}$.

It is convenient for pedagogical purposes to split V_L+V_I into two parts. V_L is the potential energy for the perfect lattice, and V_I arises from the point imperfection. Thus, in the case of the F center, V_I is due to the negative-ion vacancy. Mott and Littleton considered this problem (M12), so one may call V_I the Mott-Littleton term. In the more conventional treatment, V_L+V_I arises from the coulombic and repulsive forces of the crystal. In treatments using the Born-Oppenheimer technique, it arises from the eigenvalue of an electronic Hamiltonian which includes all the electrons except the trapped one (B, p. 173).

First, deep traps are considered; hence, we are required to use the adiabatic approximation. Equation (3.13) gives

$$(H_L+H_e)\varphi(\mathbf{R})\chi=H_0(A)\varphi(\mathbf{R})\chi$$
$$=\varphi T\chi+(t+V_L+V_I+V_e)\varphi\chi, \quad (5.2)$$

$$H_1(A)=H-H_0(A). \quad (5.2a)$$

There is a striking contrast between our problem and a photon field. The splitting in electrodynamics is done purely classically and in an elegant manner. Here we have had to introduce quantum mechanical operators and further assume an approximate wave function. How does one split H_0 into two parts?

Following the development in Sec. 3, we expand about $\mathbf{R}_n$ defined by

$$\left\{\frac{\partial}{\partial X_\alpha(k)}[\Delta\epsilon_n+V_L+V_I]\right\}_{\mathbf{R}_n}=0. \quad (5.3)$$

$\Delta\epsilon_n$ is an eigenvalue of $t+V_e(\mathbf{R},\mathbf{r})$. We split $H_0(A)$ by writing

$$H_L=T'+V_L+V_I+\Delta\epsilon_n(\mathbf{R})-\Delta\epsilon_n(\mathbf{R}_n) \quad (5.4)$$

and

$$H_e=t+V_e. \quad (5.5)$$

T' means that T only operates on χ. We note that

$$H_0\neq H_L+H_e. \quad (5.6)$$

When the sum operates on $\varphi\chi$ an extra function $\Delta\epsilon_n(\mathbf{R})-\Delta\epsilon_n(\mathbf{R}_n)$ appears. H_L includes approximately the cohesive energy of the distorted crystal.‖

The modes defined by (5.3) and (5.4) depend on the state of the electron due to the presence of $\Delta\epsilon_n$. Using the foregoing development, we may understand the various simplifications that have been made.

(1) The phenomenological approach (a) ignores the presence of V_I and (b) makes a crude approximation of the actual Hamiltonian by assuming that the equation

$$\{H_L-V_I-\Delta\epsilon_n(\mathbf{R})+\Delta\epsilon_n(\mathbf{R}_n)\}\chi=\epsilon_v\chi \quad (5.7)$$

gives pure longitudinal and pure transverse optical modes. It further assumes that the lower branches are pure acoustical.

These extreme simplifications are inadequate for some problems involving normal modes in alkali halides. The interaction between the trapped electron and the phonons is obtained by adding $\Delta\epsilon_n$ to (5.7); namely, one requires solutions of the following equation:

$$\{H_L-V_I\}\chi=\epsilon_v\chi. \quad (5.8)$$

In principle the inclusion of $\Delta\epsilon_n$ leads to no difficulties, and the method developed in Sec. 3 can be employed. Meyer has developed a perturbation scheme using V_e. This, however, leads to the omission of some terms. In Appendix II we compare both methods.

(2) The local modes approach: One assumes that the eigenequation

$$H_L\chi=\epsilon_v\chi \quad (5.9)$$

gives two types of modes, q_i, which are local, and q_c, which pertain to the perfect crystal. For this approach to be valid, $\Delta\epsilon_n$ must be independent of the q_c. If $\Delta\epsilon_n$ is only a function of a single localized mode then we have the CC model. The $\Delta\epsilon$'s can again be used to calculate the effect of the trapped electron on the modes. *A priori*, there is no clear reason for believing that $\Delta\epsilon_n$ will be independent of the q_c's. Hence, there is no apparent reason to assume that $\Delta\epsilon_n$ will not be a function of the q_c's. Only extremely complex calculations, not available at present, could tell the exact form of $\Delta\epsilon(q_i,q_c)$. One may resolve this problem somewhat by obtaining equations for the shape of absorption and emission bands in solids with adjustable parameters. These parameters can be evaluated by analysis of the experimental data. This, in turn, will give basic information regarding the relations among $\Delta\epsilon_n$, q_c, and q_i. Part II of this paper is devoted to this problem.

(b) Free Electron

Consider next a free electron in a metal or semiconductor. Two arguments suggest that the Born-Oppenheimer technique should not be used. First, consider the effect a free electron at $\mathbf{r}$ has on the modes. Its behavior is described by means of a wave packet

‖ Approximately because we use $\mathbf{R}_n$ defined by Eq. (5.3) instead of the values of $\mathbf{R}$ determined by minimizing V_L or V_L+V_I.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

made from Bloch functions. The electron proceeds in a straight line for about 200 A and then gets reflected through an angle. If the electron has the energy of 3 ev (approximately the Fermi level of a metal) it collides with the lattice every 2×10^{-14} sec. It is not confined to a local region, but wanders in random fashion, so that an average effect over a period of vibration is meaningless. Analytically, for a pure Bloch function, ϵ_j equals zero, so such an electron does not exert a force on the normal modes as shown below. Therefore, one splits the system's Hamiltonian as follows ($V_I=0$)

$$H_L=T+V_L(\mathbf{R}) \tag{5.10a}$$

$$H_e=t+V_e(\mathbf{R}_L) \tag{5.10b}$$

and

$$H_1=V_e(\mathbf{R})-V_e(\mathbf{R}_L) \tag{5.10c}$$

to obtain the usual theory (W4-Chap. IX). The normal modes about $\mathbf{R}_L$ are determined from V_L. Here another assumption is made, that one should use $\varphi(\mathbf{R}_L)$ rather than $\varphi(\mathbf{R})$. The continuously readjusting wave function requires a type of kinetic energy, and one suspects that the loss in potential energy does not compensate for this gain (M4).

The distinction between the static and the adiabatic approximation is academic, since, as Haug (H2) has shown, that first-order perturbation theory gives the same results for both. The treatment assumes that one does not use the Born-Oppenheimer technique for the free electron—that is, that H_L determines the q's. This requires that $\epsilon_j(\mathbf{R}_L)=0$. Proof of this relation for a Bloch electron is now given.

The interaction between an electron and the lattice in the conduction band has the form

$$H_1=\textstyle\sum_j a_j\Lambda_j, \tag{5.11}$$

where the Λ_j's are determined from (5.10a). To a sufficient approximation we write

$$a_j=\text{const}\,\exp(i\mathbf{k}_j\cdot\mathbf{r}) \tag{5.12}$$

where $\mathbf{k}_j$ is the wave-number vector of the jth mode. The Bloch function for a conduction electron may be written in the form

$$\varphi_p=u_p(\mathbf{r})\,\exp(i\mathbf{p}\cdot\mathbf{r}) \tag{5.13}$$

where $\mathbf{p}$ is the momentum vector and u_p has the periodicity of the lattice. Hence from (3.16) it follows that

$$\epsilon_j=\langle\varphi_p|\,a_j\,|\varphi_p\rangle=\text{const}\int u_p{}^2\exp(i\mathbf{k}_j\cdot\mathbf{r})d\tau=0 \tag{5.14}$$

since u_p is periodic. This shows that the Λ_j's describe the true modes.¶ This argument also applies to nonpolar semiconductors.

This proof as it stands should apply to polar crystals as well. Detailed considerations show that this development is incomplete. This follows because some ϵ_{jj} are zero in spite of the fact that $\Delta\epsilon_n(\mathbf{R})$ is not small. For this case H_1 has a major influence on the q's, and the presence of the electron may generate a new set of modes. Here the problem does not correspond to the one in electrodynamics.

Applying this approach, Eq. (5.10), to deep traps where the electron's kinetic energy is large, (5.14) does not hold, and exceptionally large values of the interaction are obtained, simply because the q's derived from (5.10) are not the appropriate ones. These χ's form an orthogonal set, although a force $-\epsilon_j(\mathbf{R}_L)$ remains on the nuclei at $\Lambda_j=0$.

(c) Comparison with Quantum Electrodynamics

Coming now to some general comparisons between quantum electrodynamics and the theory developed here, no equation in quantum electrodynamics corresponds to (3.4a). First, consider case b where all the $\epsilon_j(\mathbf{R}_L)$'s are zero and the Λ's are the true modes. Hence H_{rad} corresponds to H_L, H_{el} corresponds to H_e, and H_{int} corresponds to H_1. The frequency and wavelength of the photon modes are determined for free space and are not influenced by the electrons, although their number is determined in part by electronic transitions. The number of phonons is also determined by the electronic state, if we use time dependent perturbation theory, although the q's are obtained from H_L. The polaron problem in semiconductors and metals corresponds to the self-energy problem in electrodynamics. This is also true for many optical problems in valence semiconductors, i.e., such as the band-to-band transition in germanium, silicon, and diamond. Here the normal modes do indeed behave in a manner similar to photons, and the word "phonons" is most appropriate.

This is to be contrasted to the situation where method (a) should be used. The term which mixes the electronic and normal mode wave functions or causes transition is $H_1(A)$ of (5.2a). The ϵ_j's and ϵ_{jj}'s have no counterpart in quantum electrodynamics. They generate a new set of q's for every electron level and cause an essential difference between the two fields. Their presence makes multiphonon processes natural in some solids. In this case, optical problems are similar to those occurring in diatomic molecules (Part II). The similarity between phonons and photons breaks down. If one applies method (b) to polar crystals, there is no simple way to separate the energy associated with the poor choice of the modes and that caused by $H_1(S)$.

Thus we have two types of problems in phonon-dynamics. In one the ϵ's can be ignored and use can be made of the tools of modern electrodynamics. In the other, we are in a different domain, where the ele-

¶ One may arrive at the expression for ϵ_j in a more rigorous manner. In general $\langle\varphi_q|H_1|\varphi_p\rangle\neq0$ only if $\mathbf{p}-\mathbf{q}\pm\mathbf{k}=0$ [W4, Eq. (9.3.3) with $\mathbf{g}_b=0$]. Here $\mathbf{p}=\mathbf{q}$; hence, for $\mathbf{k}_j\neq0$, $\epsilon_j=0$.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

mentary concepts of electrodynamics do not serve as a guide. The first situation is in a sense a "special" case.

6. CONCLUSIONS OF PART I

1. By the use of the Born-Oppenheimer method, we have shown what effects electron transitions have on the normal modes of vibration. The problem has been formulated in a new manner to show the interrelation of the various terms. The two major effects are (a) a shift in the equilibrium position of the normal modes—first-order effect; (b) a change in the frequencies of vibration—second-order effect. The analytical reason for these shifts is shown and related to the change in the electronic binding energies of traps. The relations which are derived from the microscopic Hamiltonian are completely general. It was indicated why one would not expect the calculations made by Huang and Rhys, by Pekar, and by Lax to apply to an F center. To obtain an accurate theoretical expression, greater care must be taken in the use of various expansions.

2. Various activation energies are defined and general relations obtained between them. These impose restrictions on the interpretation of some experiments done on color centers.

3. A comparison has been made of quantum electrodynamics and phonondynamics. Terms not present in the former arise. They account for multiphonon processes. These additional terms play a very important role in the shape of optical absorption and emission bands. They may enter into the polaron problem. The electron in the conduction band of some polar solids may create new local modes about itself. These local modes do not appear in electrodynamics, hence the polaron problem in some crystals is not equivalent to problems in electrodynamics. These terms make it possible for a hole to be self-trapped, as indicated by the experiments of Castner and Kanzig (C1). To understand the stability of some of the new types of centers, the molecular color centers, we must examine the phonon-electron Hamiltonian discussed here in fuller detail.

Part II—Broadening of Absorption Lines Due to Multiphonon Processes

7. INTRODUCTION TO PART II

In the following sections the theory developed in Part I is applied to the broadening of absorption bands in solids. We only consider transitions between two bound states. The breadth of a line may be due to several causes.

(a) One is the Heisenberg principle which requires an uncertainty in the energy of a level, provided its lifetime is finite. An alternate way of looking at this problem is to use higher-order terms in the standard time-dependent perturbation theory (H3, p. 181). This explains the breadth of the hydrogenic lines at low pressure. The breadth of lines due to this effect should be very roughly of the order of that observed in atoms.

TABLE II. Breadth of absorption lines in ev.

Atomic	
One of the H_α lines of hydrogen	3×10^{-6}
A very wide atomic line of CuI ($^4F_{\frac{5}{2}}-{}^4D_{\frac{3}{2}}$ at 4540 A)	6×10^{-4}
Solids	
The 278 cm^{-1} line of B doped Si	7×10^{-4}
The F center breadth in KCl at 0°K	0.16
The F' center breadth in KCl	about 1
A narrow band in LiF (P5)	0.01
Energy of one optical phonon	0.01–0.08

(b) Broadening is also caused by the interaction between centers. We must assume that there is more than one trap in a solid so that (5.1), which applies only to an individual imperfection, does not contain the whole Hamiltonian. Their sum does not form the total Hamiltonian, for one must consider interactions between centers. Without these additional terms, ϵ_u represents a highly degenerate level, since any imperfection can be excited. With the interaction terms, the degeneracy of ϵ_u is removed, and the excited states spread out into a narrow band. Since no detailed calculation has been made, it is not known whether the degeneracy of ϵ_u can be completely removed. If a crystal has 5×10^{16} centers per cc, the mean separation is 270 A. This type of interaction must be small in polar solids, since 270 A is large compared to the spread of the electron wave functions. For example, the excited state of the F center in KCl does not extend beyond 13 A (Smith, S6). The interaction may not be small if the imperfection is such that the spread in the wave function is of the order of the distance between centers.

(c) Although this effect may make a small contribution to the broadening of an absorption line in solids, our interest is in the fact that during an optical transition no strict selection rules apply to the v_j's. If $\epsilon_j=\epsilon_{jj}=0$, strict rules would apply, and this effect would not make a contribution to the problem.

Effects (a) and (b) are essentially temperature independent, while (c) depends strongly on the temperature.

In view of the complexity of the problem, equations are derived with parameters which are to be evaluated experimentally. This is the first necessary step, since at present our knowledge is far too meager to attempt exact *a priori* calculations. Some experimental facts will be presented before embarking on the theoretical details.

The width of lines varies over a large range. Some typical values are given in Table II. Thus the width of some lines in solids approaches the broadest found in atomic spectra. In general, lines in solids are considered broader, however. In polar crystals some bands are very much wider, and the absorption or emission of a

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

single phonon will not explain the phenomenon. The marked temperature dependence of the shape of some lines strongly suggests that (c) can dominate the others. Detailed studies (M5, K4) support this conclusion.

The theory developed here does not apply to the F' band since the upper state is not bound. It is not completely clear why some bands in polar materials are extremely narrow. One may simply assume that certain parameters, to be derived, are very small. On the other hand, more complex concepts may be involved, such as the breakdown of the Born-Oppenheimer technique; further studies are required. The theory developed applies to bands whose width is approximately equal to the F center and is temperature dependent.

Experimentally, the absorption, α, is measured as a function of the wavelength or the photon energy ϵ (in ev). The observed $\alpha(\epsilon)$'s are not simple analytical curves; hence a way must be found to characterize them. We may in general define: ϵ_m, the value of ϵ where α has its maximum value, α_m; ϵ_r is where α has half its maximum value on the red side of ϵ_m, and ϵ_v is the corresponding point on the violet. Further useful definitions follow:

the moments— $$M_n=\int_0^\infty \alpha(\epsilon)\epsilon^n d\epsilon \tag{7.1}$$

the average— $$\bar{\epsilon}=M_1/M_0 \tag{7.2}$$

$$m^2=\frac{1}{M_0}\int_0^\infty (\epsilon-\bar{\epsilon})^2\alpha(\epsilon)d\epsilon \tag{7.3}$$

and the half-width— $$H=\epsilon_v-\epsilon_r. \tag{7.4}$$

In general,

$$m^2=(M_2/M_0)-\bar{\epsilon}^2. \tag{7.5}$$

The absorption is due to upward transitions induced by the vector potential of the light which falls on the crystal. To relate the measured absorption at photon energy ϵ to the transitions, four factors must be considered.

(1) The vector potential at the imperfection has to be related to the intensity of light falling on the crystal. Lorentz suggested a way of doing this. It is not known whether his theory applies rigorously to traps in solids and deviations from his simple equation would not be surprising. In any case the Poynting vector (the light intensity) is proportional to the square of the vector potential at the imperfection.

(2) The second factor is due to the φ's which give matrix elements. The expression may be simplified by the introduction of an oscillator strength, f. Classically, the second item contains no unknown since f equals unity. Quantum mechanically, the problem requires exact wave functions, which at present are unknown. Most probably, we shall have to introduce the f factor empirically for some time to come.

(3) The imperfection concentration.

(4) Finally, one must consider the shape factor. This can be done by multiplying the usual atomic absorption equation, item (2), by a function $G_n(\epsilon)$. G_n arises from items (a), (b), and (c) considered previously. In all cases studied so far, G_n has the property $\int G_n(\epsilon)d\epsilon \cong 1$.

The measured absorption at ϵ is the product of the above four factors. If one could make exact calculations, the impurity concentration could be treated as an unknown and computed from a single point, ϵ. Even without an exact calculation, one might evaluate the factor associated with (1) and (2) from experimental data and obtain the concentration with the use of a single constant. This would require a detailed knowledge of G_n, which is lacking. The G_n can be eliminated by an integration over ϵ. This procedure is impractical in many cases, for one cannot measure the absorption at the wings of the band (due to overlapping from other absorption bands). The integration technique is also extremely lengthy. Many years ago Smakula suggested replacing $\int \alpha(\epsilon)d\epsilon$ by a product of α_m times H. This requires the introduction of a constant a_s such that $a_s G_m H=1$, where G_m is the peak value of G.

To illustrate this procedure, let us assume that G_n has the triangular shape, i.e.,

$$\begin{aligned} G_n &= 0 && \text{for} \quad \epsilon > a \\ &= (1/H)(a-\epsilon/a-\epsilon_m) && \text{for} \quad a > \epsilon > \epsilon_m \\ &= (1/H)(\epsilon-b/\epsilon_m-b) && \text{for} \quad \epsilon_m > \epsilon > b \\ &= 0 && \text{for} \quad b > \epsilon. \end{aligned} \tag{7.6}$$

The area under G_n is unity. The integration over the frequency equals $G_m H$ and $a_s=1$ for the shape given by (7.6). Hence, the product $G_m H$ or $\alpha_m H$ eliminates the fourth factor. The advantage of this procedure is that the temperature dependence of H seems to be eliminated.**

The essence of Smakula's famous equation is the substitution of $G_m H$ for $\int G(\epsilon)d\epsilon$, not the exact form of $G_n(\epsilon)$. Therefore we refer to the general equation by his name and call a_s Smakula's constant. In this equation a_s is always multiplied by two factors arising from (1) and (2); hence the actual value of a_s is of little experimental importance. Nevertheless, a_s is a measure of the shape of an absorption band, as are the variables defined in Eq. (7.1) to (7.5).

It is useful to determine H, m, and a_s for several types of $G_n(\epsilon)$, namely,

The Gaussian

$$G_n=\frac{a}{\sqrt{\pi}}\exp\{-a^2(\epsilon-\epsilon_m)^2\}. \tag{7.7}$$

** "Seems to be," since there is practically no experimental data to support or refute this assumption.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

Hence,

$$H(G)=\frac{1}{a}[4\ln 2]^{\frac{1}{2}}. \tag{7.7a}$$

Further

$$m^2=(1/2a^2)=(1/8\ln 2)H^2=(1/5.545)H^2 \tag{7.7b}$$

and

$$a_s(G)=(1/G_mH)=[\pi/4\ln 2]^{\frac{1}{2}}=1.0645. \tag{7.7c}$$

The Lorentzian

$$G_n=\frac{b}{\pi}\frac{1}{[1+b^2(\epsilon-\epsilon_m)^2]} \tag{7.8}$$

Now

$$H(L)=(2/b) \tag{7.9a}$$

and

$$a_s(L)=(\pi/2)=1.5708. \tag{7.9b}$$

M_2 does not exist for (7.8).

The Double Gaussian

G_n might be made of two Gaussians, one for the violet side with $a=a_v$ and another for the red side, with $a=a_r$. Then

$$G_m=(2/\sqrt{\pi})[(1/a_r)+(1/a_v)]^{-1}, \tag{7.10}$$

$$H(DG)=(\ln 2)^{\frac{1}{2}}[(1/a_r)+(1/a_v)], \tag{7.11a}$$

$$m^2=\tfrac{1}{2}[(1/a_r)+(1/a_v)]^{-1}\times[(1/a_r^3)+(1/a_v^3)], \tag{7.11b}$$

and

$$a_s(DG)=a_s(G)=1.0645. \tag{7.11c}$$

Various attempts have been made to describe the shape of the absorption band associated with F centers and all three types have been considered. One of the most successful is the double Gaussian. For this curve a_s is the same as for the single Gaussian.

The Pekarian

This curve is discussed in Sec. 9 and Appendix IV. To obtain $a_s(P)$, numerical integration techniques were used, since the analytic form is complex. It actually depends on H. However, unless the half-width is very small,

$$m^2=H^2/5.57 \tag{7.12a}$$

and

$$a_s(P)=1.07, \tag{7.12b}$$

which agrees with Eqs. (7.7b) and (7.7c). A Pekarian curve resembles a double Gaussian, so that the similarity of $a_s(P)$, $a_s(G)$, and $a_s(DG)$ is not surprising.

In Smakula's original derivation $a_s(L)$ was used. In recent years experimental studies indicate that the Gaussians or double Gaussians give better agreement with the observed curves. Dexter has suggested that one might replace $a_s(L)$ by $a_s(G)$. There are good arguments for this change. The absorption shapes, however, are not true Gaussians, and the suggestion simply means that we replace one empirical constant with another without gaining any fundamental information. There is some evidence that the band shape is Pekarian.

Since in Smakula's equation a_s is always multiplied by two other factors, the actual value of a_s is of little experimental importance. The danger in making the change is that confusion will result. The day may come when one will know the true form of G_n and have a correct expression for a_s. As this day has not yet arrived, the standard form of Smakula's equation is preferable.

8. GENERAL FORMULATION OF ABSORPTION PROBLEM

In this section the absorption problem is formulated in general. The next two sections discuss the shape and moments. Since deep traps are to be considered, the adiabatic approximation must be used. The electron-phonon interaction term, i.e., (3.13b), will be ignored. The Hamiltonians of the solid can therefore be written in the forms: for the ground state

$$H_g=\epsilon_g(\theta)+T_g'+V_g \tag{8.1}$$

or

$$E=\epsilon_g(\theta)+\textstyle\sum_j \hbar\omega_j(g)(v_j+\frac{1}{2}), \tag{8.1a}$$

while for the excited state

$$H_u=\epsilon_u(\theta)+T_u'+V_u \tag{8.2}$$

or

$$E'=\epsilon_u(\theta)+\textstyle\sum_j \hbar\omega_j(u)(v_j'+\frac{1}{2}), \tag{8.2a}$$

where v_j is the vibrational quantum number of the jth mode when the electron is in the ground state, while v_j' corresponds to the upper state; ϵ_g and ϵ_u are evaluated at $\mathbf{R}_g$ and $\mathbf{R}_u$, (3.19a) and (3.19b) $H_{g,u}$ is an operator, while E is the total energy, (3.14a). V_g and V_u are $\epsilon_g(\mathbf{R})-\epsilon_g(\mathbf{R}_g)$ and $\epsilon_u(\mathbf{R})-\epsilon_u(\mathbf{R}_u)$, respectively [see Eq. (3.4)].

Effects of thermal expansions are not included in the theory of Part I. The thermal expansion of the crystal arises from the dependence of the vibrational frequencies on the volume (Slater, S5, p. 199). In principle, the thermal expansion may influence all the variables. It is believed that the primary effect of the temperature (θ) is on ϵ_g and ϵ_u.

To arrive at an expression for the absorption coefficient, we start with the standard equation for the energy absorbed by an atom when light falls on it (2 of Sec. 7). This expression must be corrected for the presence of the medium and for various broadening effects, as discussed in the last section (1 and 4). Since Dirac's delta functions are used, special care must be employed. First-order perturbation theory gives the following expression for the energy absorbed by an atom in a

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

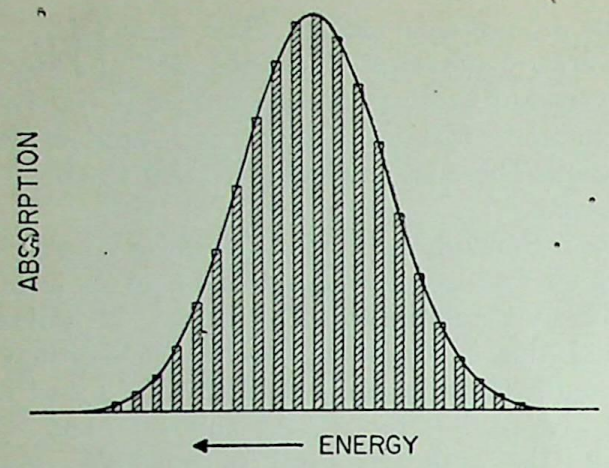


FIG. 4. Schematic diagram showing the broadening due to the emission or the absorption of phonons.

vacuum (H3, p. 180):

$$s_{\text{atom}}=\frac{4\pi^2}{3}\frac{e^2}{\hbar c}\nu|\{\langle\Psi'(A)|\mathbf{r}|\Psi(A)\rangle\}|^2 I_\nu, \qquad (8.3)$$

where ν is the angular frequency of the absorbed light††; $\Psi'(A)$ and $\Psi(A)$ are the total wave functions for the ground and excited states, Eq. (3.11); and I_ν is the intensity of the light per unit angular frequency interval, i.e., the energy of the light which falls on the crystals in the angular frequency interval $\Delta\nu$ is $I_\nu\Delta\nu$.

In a medium, the following modifications must be made.

(1a) If n_0 is the high-frequency dielectric constant, the intensity I_ν (Poynting vector) is $(cn_0F^2/4\pi)$. Since the transition is caused by F^2, (8.3) must be divided by n_0.§§

(1b) The space average field must be replaced by the local field. Since they are proportional to each other, we introduce a constant bn_0. For a Lorentz local field $b=(1/9n_0)(2+n_0^2)^2$ (B, p. 100).

(2) As discussed, lines have a finite width.

(1a) and (1b) correspond to 1 of Sec. 7. (2) corresponds to the broadening due to the uncertainty principle and the interaction between neighboring centers. If (8.3) is multiplied by b, the effects of (1a) and (1b) are included.

(2) requires the introduction of the function $g(\nu)$ with the following properties; it peaks at

$$\nu=[E'-E]/\hbar \qquad (8.4)$$

and

$$\int_0^\infty g(\nu)d\nu=1. \qquad (8.5)$$

The g does not include the effects of the ϵ_j's and ϵ_{jj}'s and is not the same as G_n. Figure 4 illustrates the problem. There is an over-all absorption band and a fine structure. We may expect to observe the fine structure only if the half-widths of the g's are less than the energy of one phonon (i.e., $\hbar\omega$). Figure 4 assumes that one is dealing with a single phonon frequency. When a center interacts with phonons of several frequencies, i.e., if there is dispersion, the ϵ_j's and ϵ_{jj}'s can cause a broadening of the "lines" shown in Fig. 4. This problem has been discussed by Krivoglaz and Pekar (K5). There seems to be no simple way to separate this effect from those discussed in Sec. 7. This effect is not considered further here.

We now digress slightly to indicate how the trap-trap interaction produces a broadening. The breadth of an electron-spin resonance line also arises from two effects: one is because of the interaction between the spins (spin-spin), and the other is because of the interaction of the electron with the magnetic moments of the surrounding nuclei (hyperfine). The hyperfine interaction often has a much larger effect. The problem is thus similar to the one considered here. Strangely enough, we have a much better understanding of the broadening problem in the newer field (electron-spin resonance).

For this development the solid is considered to be a large molecule of unit volume. There will therefore be N imperfections. We ignore the lattice vibrations, hence denote by $\mathbf{R}_j$ the rest position of the jth imperfection, and by $\mathbf{r}_j$ the position of its electron (relative to $\mathbf{R}_j$). Again two bound states per trap are assumed. The total ground state eigenfunction has the following form in the zeroth approximation, where the interactions between imperfections are ignored,

$$\Phi_0(g)=\Pi^j\varphi(\mathbf{R}_j+\mathbf{r}_j). \qquad (8.6)$$

(8.6) uses the Hartree approximation (S2, p. 234), while the excited state is of the form

$$\Phi_0(u_1k)=\varphi(\mathbf{R}_1+\mathbf{r}_1)\cdots\varphi(\mathbf{R}_{k-1}+\mathbf{r}_{k-1})\times\varphi'(\mathbf{R}_k+\mathbf{r}_k)\varphi(\mathbf{R}_{k+1}+\mathbf{r}_{k+1})\cdots\varphi(\mathbf{R}_N+\mathbf{r}_N). \qquad (8.7)$$

It is N-fold degenerate, since any imperfection can be excited.

We next consider two perturbations: (1) due to the optical waves and (2) due to the imperfection-imperfection interaction. The first has the form [H3, Eq. (13.7)],

$$\sum_j(e/mc)\mathbf{p}_j(e)\cdot\mathbf{A}(\mathbf{R}_j+\mathbf{r}_j), \qquad (8.8)$$

where $\mathbf{p}_j(e)$ is the electron's momentum and $\mathbf{A}$ is the vector potential. If one combines (8.6), (8.7), and (8.8) with the usual theory of absorption (H3, p. 143 and p. 179), the following transition matrix is obtained,

$$|\langle\Phi_0(u,k)|\mathbf{R}_{\text{eff}}|\Phi_0(g)\rangle|^2 \qquad (8.9)$$

where

$$\mathbf{R}_{\text{eff}}=\sum_j\mathbf{r}_j\exp(-i\boldsymbol{\kappa}\cdot\mathbf{R}_j) \qquad (8.9a)$$

†† The angular frequency of an electromagnetic wave is denoted by ν, while the angular frequency of a phonon is ω. The use of ν as an angular frequency is not the usual practice, but it has been used in one standard treatment (H3).

§§ The permeability is assumed to be unity (A1, p. 188). F is the electric field strength.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

and κ is the wave vector (ν/c) of the exciting light. The total effect is a sum of (8.9) over all the upper states. In this approximation (8.9) is just $N|\langle\varphi'|\mathbf{r}|\varphi\rangle|^2$, as one would expect.

The second perturbation produces changes in (8.6) and (8.7). We first apply the trap-trap interaction and then use (8.8). Since $\Phi_0(g)$ is nondegenerate, its energy level will be slightly shifted. The perturbation, however, will affect $\Phi_0(u,k)$ more profoundly. It removes some or all of the degeneracy and the upper level will fan out (Fig. 5). The various eigenfunctions are denoted by $\Phi_1(u,k)$. One may no longer replace $\mathbf{R}_{\text{eff}}$ by $\mathbf{r}$, since the phase factor is of utmost importance; hence, the matrix element becomes

$$\langle\Phi_1(u,k)|\mathbf{R}_{\text{eff}}|\Phi_1(g)\rangle. \tag{8.10}$$

Many of the cross terms in the square of the matrix cancel due to the random distribution of the $\mathbf{R}_{\text{eff}}$. The old theory of spectroscopic stability (V1, p. 137) gives some useful information. It can be applied directly to (8.10) to show that

$$\sum_k^N |\langle\Phi_0(u,k)|\mathbf{R}_{\text{eff}}|\Phi_0(g)\rangle|^2 = N|\langle\varphi'|\mathbf{r}|\varphi\rangle|^2$$

$$= \sum_k^N |\langle\Phi_1(u,k)|\mathbf{R}_{\text{eff}}|\Phi_1(g)\rangle|^2. \tag{8.11}$$

The perturbation is assumed to be small enough that the $\Phi_1(u,k)$'s can be expanded in terms of the $\Phi_0(u,k)$'s.

The next step is to define $g(\nu)$ by the relation

$$\sum_{\nu-\frac{1}{2}\Delta\nu}^{\nu+\frac{1}{2}\Delta\nu} |\langle\Phi_1(u,k)|\mathbf{R}_{\text{eff}}|\Phi_1(g)\rangle|^2$$

$$= Ng(\nu)|\langle\varphi'|\mathbf{r}|\varphi\rangle|^2\Delta\nu. \tag{8.12}$$

The summation is restricted to states where the energy difference lies between $\hbar\nu\pm\frac{1}{2}\hbar\Delta\nu$. The left-hand expression is simply the sum over all the possible transitions, while the right-hand side defines $g(\nu)$. The theorem of spectroscopic stability assures that $\int_0^\infty g d\nu = 1$ as required by (8.5).

The development is quite formal and gives no information regarding the shape and width of g. These are the required steps, however, to introduce the factor properly into the theory of optical absorption. It shows that the upper level widens.

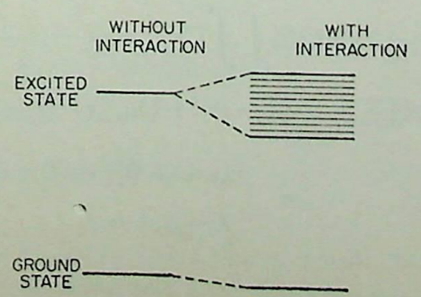


FIG. 5. Schematic diagram showing the effect of the interaction between centers.

This development indicates why one would not expect sharp lines in the absorption spectra of solids, except when one "imperfection" is very well shielded from another. Hence the absence of sharp lines in the spectroscopy of solids is to be expected (see the edge emission of CdS. K2a, p. 117). This factor also accounts for the fact that the F center absorption does not have a fine structure. Unfortunately, we are unable to estimate the width of g, or to separate this broadening from the one caused by dispersion (K5).

It is assumed that g is wider than $\hbar\omega$, where ω is an effective phonon angular frequency. The g function is next replaced by the Dirac δ function. To obtain the moments, we integrate over many such functions. For the shape calculation, more care has to be used. On introducing the δ function into (8.3) and recalling items (1a) and (1b), we obtain

$$s(\nu) = b(4\pi^2/3)(e^2/\hbar c)\nu|\{\langle\Psi'(A)|\mathbf{r}|\Psi(A)\rangle\}|^2 \times\delta\{\nu-(1/\hbar)(E'-E)\}I_\nu. \tag{8.13}$$

The dimensions of s_{atomic} and $s(\nu)$ are not the same. s_{atomic} has the dimensions of ml^2t^{-3} (m for mass, l for length, and t for time). This arises because I_ν has the dimension of mt^{-2}, not energy per l^2t. Equation (8.3) assumes implicitly a δ function (S2, p. 215). In (8.13) we have written the δ function explicitly; hence $s(\nu)$ has the dimensions of ml^2t^{-2}.

The absorption cross section per imperfection in cm^2 is defined by

$$\frac{\text{energy absorbed per unit time in interval } \Delta\nu}{\text{energy falling on imperfection per unit time for unit area in interval } \Delta\nu}$$

or

$$\sigma(\nu) = b(4\pi^2/3)(\nu e^2/\hbar c)|\{\langle\Psi'(A)|\mathbf{r}|\Psi(A)\rangle\}|^2 \times\delta\{\nu-(1/\hbar)(E'-E)\}. \tag{8.13a}$$

This cross section must be summed over all the point imperfections in unit volume. Before summing (8.13a) we must consider the probability of finding a given imperfection with energy E. This is given by (B, p. 178)

$$P(v_t) = \Pi^i p(v_j), \tag{8.14}$$

where

$$p(v_j) = e^{-v_j\beta}(1-e^{-\beta_j}) = 2e^{-(v_j+\frac{1}{2})\beta_j}\sinh\tfrac{1}{2}\beta_j \tag{8.14a}$$

and

$$\beta_j = \hbar\omega_j(g)/k\theta. \tag{8.14b}$$

The time average absorption constant is $\sum_{v,v'}P(v_t)\cdot\sigma(\nu)$. To obtain the actual measured absorption coefficient, the δ function must be eliminated. This can be done by the integration

$$\alpha_\omega = \frac{1}{\omega}N\int_{\nu-\frac{1}{2}\omega}^{\nu+\frac{1}{2}\omega}\sum_{v,v'}P(v_t)\sigma(x)dx. \tag{8.15}$$

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

On averaging $\sigma(\nu)$, the error of replacing $\hat{g}$ by δ is eliminated. For the calculation of the moments, one need not use (8.15) but simply integrate $\sum_{v,v'} P(v_t)\sigma(\nu)$ over many narrow "lines."

To carry the solution further, the delta functions must be expressed in terms of integrals. ν appears in the δ function and as a multiplier of this function. If the oscillator strength‖ ‖ is introduced (S4),

$$f(\mathbf{R}) = (2/3)(m/\hbar)\nu |\langle \varphi'(\mathbf{R}) | \mathbf{r} | \varphi(\mathbf{R}) \rangle|^2. \quad (8.16)$$

α_ω takes the following form

$$\alpha_\omega(\nu) = \frac{1}{\omega} bN \int_{\nu-\frac{1}{2}\omega}^{\nu+\frac{1}{2}\omega} d\nu \left[\frac{2e^2\pi^2}{cm} \sum_{v,v'} P(v_t) |\{\chi'|f^{\frac{1}{2}}|\chi\}|^2 \times \delta\left\{ \nu - \frac{1}{\hbar}(E'-E) \right\} \right], \quad (8.17)$$

which nicely hides the frequency dependence of the term in front of the δ function. The mass of the electron (m) is introduced by using (8.16). It can be considered as an adjustable parameter provided we have a knowledge regarding the behavior of f. The next step is to forget about f's dependence on ν and $\mathbf{R}$ and write¶¶

$$\alpha_\omega(\nu) = \frac{2\pi^2 be^2}{c\omega m} fN \int_{\nu-\frac{1}{2}\omega}^{\nu+\frac{1}{2}\omega} G(x)dx \quad (8.18)$$

$$G_n(\nu) = \sum_{v,v'} P(v) |\{\chi'|\chi\}|^2 \delta\{\nu - (1/\hbar)(E'-E)\}. \quad (8.18a)$$

In Sec. 10 we shall show that G_n is normalized. In (8.18) the shape and moments only depend on the χ's. The assumption that f is independent of $\mathbf{R}$ is known as the Condon approximation. Meyer has eliminated this restriction. There is no good theoretical justification for the steps taken in obtaining (8.18). The frequency spread may be large over the bands, and one can hardly replace ν by an "average" value. In emission, a ν^3 appears before the δ function, and it certainly cannot be ignored. For instance, from the data on the emission from the F center in KCl at 77°K (D3) we know that $\epsilon_r(em) = 1.08$ ev and $\epsilon_v(em) = 1.35$ ev. Since $(1.25)^3 = 2$, the frequency factor could lead to an experimentally detectable skewness. One can assume that f is a function of the q's and of ν and still solve the moments problem. The present absorption data are not good enough to justify these added complications; they are considered only in Appendix V.

The theoretical problem is to know whether to treat $|\langle \varphi'(\mathbf{R}) | \mathbf{r} | \varphi(\mathbf{R}) \rangle|$ or $|\langle \varphi'(\mathbf{R}) | \mathbf{p}_j(e) | \varphi(\mathbf{R}) \rangle|$ as the fundamental quantity independent of the frequency. The sum rules are in terms of the f (S4, p. 503). Dirac's formulation of the radiation problem deals with matrix elements of $\mathbf{p}_j(e)$, while the elementary theory uses the matrix elements of the electron displacement. To refine the theory, this problem must first be resolved. This has not be done so far.

One cannot simply borrow theorems which hold for atoms since the phonon modes give additional degrees of freedom to the problem and their effects on the sum rules have to be explored with care.

Peierls, many years ago (S2, p. 671), considered the absorption problem when the centers have a lattice geometry. In this case the $\Phi_1(u,k)$'s have a rather simple form and strict selection rules exist for the transitions. The problem here has only an indirect relation to the one considered by Peierls. We assume a random distribution of imperfections; hence, no strict selection rules.

Lax and Burstein (L2a) have formulated the above in an alternate manner. They consider H_1 as a perturbation which broadens the Ψ's making bands out of the ground and excited states. This is an interesting and useful point of view.

Although there are limitations on (8.18a) we shall now use it to obtain the absorption coefficient when the ϵ_{jj}'s equal zero (Sec. 9), and the distribution moments in the most general case (Sec. 10). The foregoing equations apply rigorously to all the models discussed in the general introduction.

9. SHAPE OF THE BAND

Using a development of O'Rourke and Pekar, the integral $G_n(\nu)$ can be evaluated giving expressions for α. The problem, however, has to be simplified by the use of two assumptions: (1) that $\epsilon_{jj} = 0$, and (2) that an effective mean frequency exists. Use is made of Mehler's formula and various expansions of the modified Bessel function.

The delta function can be expressed in the integral form (H3, p. 66)

$$\delta(\nu) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-i\nu t} dt. \quad (9.1)$$

By substituting (8.14a) and (9.1) into (8.18a), one obtains

$$G_n(\nu) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-i(\nu-\nu_{ug})t} \Pi^j F_j(t) dt, \quad (9.2)$$

where

$$F_j = 2 \sinh \tfrac{1}{2}\beta_j \int\int [\textstyle\sum_{v_j} \exp[-(v_j+\frac{1}{2})\lambda_j] \chi_j(q_j)\chi_j(q_j')] \times [\textstyle\sum_{v_j'} \exp[-(v_j'+\frac{1}{2})\mu_j] \chi_j'(Q_j)\chi_j'(Q_j')] dq_j dq_j' \quad (9.3)$$

$$\nu_{ug} = (1/\hbar)[\epsilon_u(\theta) - \epsilon_g(\theta)] \quad (9.3a)$$

$$\lambda_j = \beta_j + i\omega_j t \quad (9.3b)$$

and

$$\mu_j = -i\omega_j t. \quad (9.3c)$$

‖ ‖ $\langle \varphi' | \mathbf{r} | \varphi \rangle$ is assumed to be real; thus $f^{\frac{1}{2}}$ is real. This allows a simplification in the notation without any limitation on the derived equations.

¶¶ Since the absorption is small, we may assume that n is independent of ν (B5).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

To obtain (9.3) we have employed (3.9), (3.24), (8.1a), and (8.2a). Also, a term has been added and subtracted in the exponent to make the next step simpler. μ_j arises from (9.1), while λ_j is a result of a combination of (9.1) and (8.14a). Finer mathematical points, such as the use of (9.1) and the rearrangement of the order of the summation will not be considered.

O'Rourke now uses Mehler's formula which can be written

$$\sum_{v_j} e^{-(v_j+\frac{1}{2})\xi}\chi_j(q_j)\chi_j(q_j') = \alpha_j(2\pi\sinh\xi)^{-\frac{1}{2}}\exp\{-\tfrac{1}{4}\alpha_j^2[(q_j+q_j')^2\tanh\tfrac{1}{2}\xi + (q_j-q_j')^2\coth\tfrac{1}{2}\xi]\}, \quad (9.4)$$

where

$$\alpha_j^2=\omega_j/\hbar \quad (9.4a)$$

(Appendix III). The substitution of (9.4) into (9.3) makes the integrand take the simple form, $\exp[-ax^2+b]$. The operation can now be carried out by the use of standard formula, giving

$$F_j=\exp\left\{\frac{-\epsilon_j^2}{\omega_j^2(\omega_j\hbar)[\coth(\frac{1}{2}\beta_j+\frac{1}{2}i\omega_j t)-\coth\frac{1}{2}i\omega_j t]}\right\}$$

$$=\exp\left\{-\left(\frac{\epsilon_j^2}{2\omega_j^3\hbar}\right)[\coth\tfrac{1}{2}\beta_j - i\sin\omega_j t - \coth\tfrac{1}{2}\beta_j\cos\omega_j t]\right\}. \quad (9.5)$$

The steps between (9.3) and (9.5) are essentially elementary, although slightly involved. The use of Mehler's formula does not require that the frequency of the two states be equal. One may obtain more general expressions for the F_j's [Vasileff (V1), or Dexter (D1)]. We want to obtain $\alpha_\omega(\nu)$, hence must obtain a rather simple expression for $\Pi_j F_j$. This is the reason for requiring that all the ϵ_{jj} vanish.

Now

$$\Pi^j F_j=\exp\left\{\sum_j -\tfrac{1}{2}\left(\frac{\epsilon_j^2}{\hbar\omega_j^3}\right)\times(\coth\tfrac{1}{2}\beta_j - i\sin\omega_j t-\coth\tfrac{1}{2}\beta_j\cos\omega_j t)\right\}. \quad (9.6)$$

The existence of the following equality (for any θ and t) must be assumed

$$\sum_j\left(\frac{\frac{1}{2}\epsilon_j^2}{\hbar\omega_j^3}\right)[\coth\tfrac{1}{2}\beta_j - i\sin\omega_j t-\coth\tfrac{1}{2}\beta_j\cos\omega_j t] = [\coth\tfrac{1}{2}\beta - i\sin\omega t-\coth\tfrac{1}{2}\beta\cos\omega t]\sum_j\frac{\frac{1}{2}\epsilon_j^2}{\hbar\omega_j^3}. \quad (9.7)$$

The relation certainly holds for the CC and Fröhlich's model. If the local modes cluster about a point between the acoustical and optical branches, (9.7) should be a good approximation. One would not expect (9.7) to hold if the important modes are a combination of several types (longitudinal optical and local) involving radically different frequencies. Combined with (8.14b), it defines the "effective" or "average" frequency ω. In Sec. 11 a method is described which can be used to determine whether or not several effective frequencies exist. For the time being, (9.7) is used. Introducing the famous Huang and Rhys factor, we write

$$G_n(\nu)=\frac{1}{2\pi}\int_{-\infty}^{\infty}dt\,e^{-i(\nu-\nu_{u0})t}\times\exp\{-S[\coth\tfrac{1}{2}\beta - i\sin\omega t-\coth\tfrac{1}{2}\beta\cos\omega t]\}, \quad (9.8)$$

where

$$S=\tfrac{1}{2}\sum_j\frac{\epsilon_j^2}{\omega_j^3\hbar}=\sum_j\frac{1}{\omega_j\hbar}\tfrac{1}{2}\omega_j^2(Q_j-q_j)^2, \quad (9.8a)$$

S is nondimensional. Equation (9.8) resembles integrals which appear in the definition of Bessel's function. To relate G_n to these functions, the following transformation is introduced:

$$\eta=(1/2)i\beta. \quad (9.9)$$

The integration variable of (9.8) is now changed to

$$x=\omega t-\eta. \quad (9.10)$$

Since one may show, with the use of (9.9) and (9.10) that

$$i\sin\omega t+\coth\tfrac{1}{2}\beta\cos\omega t=\operatorname{csch}\tfrac{1}{2}\beta\cos x, \quad (9.11)$$

Eq. (9.8) takes the form***

$$G_n(\nu)=\frac{1}{2\pi}\int_{-\infty}^{\infty}\frac{dx}{\omega}\exp\{-i(p\eta+px) - S\coth\tfrac{1}{2}\beta+S\operatorname{csch}\tfrac{1}{2}\beta\cos x\}, \quad (9.12)$$

where

$$p=(\nu-\nu_{u0})/\omega. \quad (9.13)$$

One may show that†††

$$\frac{1}{2\pi}\int_{-\infty}^{\infty}\exp\{y\cos\xi-ip\xi\}d\xi=\sum_{-\infty}^{\infty}\delta(p-k)I_p(y), \quad (9.13a)$$

where I_p is the modified Bessel function and k is an integer. Now (9.12) takes the form

$$G_n(\nu)=\frac{e^{-ip\eta-S\coth\frac{1}{2}\beta}}{\omega}\sum_{-\infty}^{\infty}\delta(k-p)I_p[S\operatorname{csch}\tfrac{1}{2}\beta]. \quad (9.14)$$

*** In (9.12) an alternate path of integration is used. This can be justified by taking contour integrals and assuming that p is an integer.

††† To establish (9.13) one must use an integral expression for I_n (J1, p. 547), and it is noted that $\delta(x)=\sum_{-\infty}^{\infty}e^{2\pi inx}$ for $-\frac{1}{2}<x<\frac{1}{2}$, etc.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

Equation (9.14) requires that p or $(1/\omega)\sum_j \omega_j(v_j'-v_j)$ be an integer for any value of v_j and v_j'. This means that $\omega=\omega_j$. Evidently Eq. (9.7), which holds for any t, can only be satisfied under this condition. Therefore, the derivation holds exclusively for this condition.

The author believes that this is the result of using the δ function in Eq. (8.13) and that in reality Eq. (9.14) will give the correct shape (envelope—see Fig. 4) even if the frequencies vary over a small range. We thus write

$$p=v_t'-v_t. \tag{9.15}$$

Equation (9.14) can be put in the alternate form

$$G_n(\nu)=\left[\frac{\bar{v}+1}{\bar{v}}\right]^{p/2}\frac{1}{\omega}\exp[-S(2\bar{v}+1)] \times\sum_k\delta(k-p)I_p\{2S[\bar{v}(\bar{v}+1)]^{\frac{1}{2}}\}, \tag{9.16}$$

where $\bar{v}=\frac{1}{2}[\coth(\beta/2)-1]$, the mean quantum number of a normal mode in the ground state.

The actually observed shape is obtained from (8.18); hence,

$$\frac{1}{\omega}\int_{\nu-\frac{1}{2}\omega}^{\nu+\frac{1}{2}\omega}G_n(\nu)d\nu=\frac{1}{\omega}\left[\frac{\bar{v}+1}{\bar{v}}\right]^{p/2} \times\exp[-S(2\bar{v}+1)]I_p\{2S[\bar{v}(\bar{v}+1)]^{\frac{1}{2}}\}, \tag{9.17}$$

where p is the integer closest to $(1/\omega)(\nu-\nu_{u0})$. ω appears in (9.17) because the unknown g function has been effectively replaced by the step function.

$$g=0 \quad \text{for} \quad \nu<\frac{1}{\hbar}(E'-E)-\tfrac{1}{2}\omega$$

$$g=\frac{1}{\omega} \quad \text{for} \quad \frac{1}{\hbar}(E'-E)-\tfrac{1}{2}\omega\leq\nu\leq\frac{1}{\hbar}(E'-E)+\tfrac{1}{2}\omega \tag{9.18}$$

$$g=0 \quad \text{for} \quad \frac{1}{\hbar}(E'-E)+\tfrac{1}{2}\omega<\nu.$$

The factor $1/\omega\exp[-S(2\bar{v}+1)]$ does not affect the frequency dependence of the absorption, i.e., the shape. It is convenient to introduce an average shape factor

$$G_A=\left[\frac{\bar{v}+1}{\bar{v}}\right]^{p/2}I_p\{2S[\bar{v}(\bar{v}+1)]^{\frac{1}{2}}\}. \tag{9.19}$$

The term $1/\omega$ is a measure of the width of the band. Larger ω's cause wider absorption bands provided that the ϵ_j's remain the same. We shall see that $\int_{-\infty}^{\infty}e^{-S(2\bar{v}+1)}G_A dp=1$ so that $\int G_n(\nu)d\nu=1$ (Appendix IV and Sec. 10).

Equation (9.19) is exact for the models considered provided (9.7) applies. It does not require that $\epsilon_j\to 0$. This is not obvious unless the derivation of O'Rourke is used. Two alternate forms of G_A will now be found using some developments originally due to Pekar (P1).

(a) Low Temperature

If $\theta\to 0$, $\bar{v}$ is a small quantity. It is useful to define $z=2S[\bar{v}(\bar{v}+1)]^{\frac{1}{2}}$ which is also a small quantity. Here $v_t\to 0$, and $p>0$; hence, our concern is for the situation where p is a positive integer. Now we expand I_p as follows (J1, p. 542):

$$I_p(z)=\frac{(\frac{1}{2}z)^p}{p!}\left[1+\frac{(\frac{1}{2}z)^2}{p+1}+\frac{(\frac{1}{2}z)^4}{2(p+2)(p+1)}+\cdots\right] \tag{9.20}$$

and obtain

$$G_A(\nu)\approx G_l(\nu)=S^p/p!. \tag{9.21}$$

This approximation improves as p increases. For large S, one may use the Stirling approximation and show that the maximum occurs at

$$p_l=S. \tag{9.22}$$

This corresponds to the angular frequency

$$\nu_l=\nu_{u0}+S\omega. \tag{9.22a}$$

Since p is an integer, S in (9.22) refers to the nearest integer value. The violet side of the band corresponds to higher values of p, where (9.21) is a better approximation.

It seems appropriate to refer to (9.21) by the name Pekarian, since Pekar was the first to employ expansion (9.20) in relation to this problem. The use of Eq. (9.21) requires the inequality,

$$z^2/4p\ll 1. \quad \text{(Pekarian Criterion)} \tag{9.23}$$

At the point of maximum absorption this is equivalent to the conditions

$$S\bar{v}\ll 1. \tag{9.24}$$

The shape of the absorption band of some F centers seems to be given by (9.21) (K4). Little is known about this type of curve. Its behavior is studied in Appendix IV.

(b) High Temperature

At high temperatures $\bar{v}$ is not small and $z\gg 1$. As the temperature increases, z approaches the value $2Sk\theta/\hbar\omega$. This suggests the use of the asymptotic expansion‡‡‡

$$I_p(z)=\frac{e^z}{(2\pi z)^{\frac{1}{2}}}\left\{1+\sum_{n=1}^{\infty}\frac{(-1)^n[4p^2-1][4p^2-3^2]\cdots[4p^2-(2n-1)^2]}{n!(8z)^n}\right\}. \tag{9.25}$$

‡‡‡ This type of expansion is discussed by Jeffreys and Jeffreys (J1). Their equation for I_p, however, is slightly in error.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

Our interests are not limited to small values of p or to the first few terms of the sum. p may have the value of 100; hence, for appropriate values of z the major contribution arises from terms with high values of n. If the series can be broken off before $r \approx p$, the simple form is obtained,

$$I_p(z) = \frac{e^z}{(2\pi z)^{\frac{1}{2}}}\left\{1 - \frac{p^2}{2z} + \frac{1}{2!}\left(\frac{p^2}{2z}\right)^2 - \frac{1}{3!}\left(\frac{p^2}{2z}\right)^3 + \cdots\right\}$$

$$= \frac{1}{(2\pi z)^{\frac{1}{2}}} \exp\left(z - \frac{p^2}{2z}\right). \tag{9.26}$$

For (9.26) to hold, one requires that $(2n-1)^2 \ll 4p^2$ for all important terms. This will be true if $p^2/2z < 1$. When $p^2/2z > 1$, the expansion (9.26) must be explored further. The highest terms of the series occur (approximately) at $n = p^2/2z$, or Eq. (9.26) requires that $4p^2 > (p^4/z^2)$. These facts suggest that criteria for the applicability of (9.26) are

$$|2z/p| > 1 \tag{9.27a}$$

and (Gaussian criteria)

$$|p| \gg 1. \tag{9.27b}$$

The high-temperature shape is obtained by substituting (9.26) in (9.19),

$$G_A(\nu) \approx G_h(\nu)$$

$$= \frac{1}{(2\pi z)^{\frac{1}{2}}} \exp\left\{z - \frac{1}{2z}p^2 + \frac{p}{2}\ln\left(1 + \frac{1}{\bar{v}}\right)\right\}. \tag{9.28}$$

Its maximum occurs at

$$p_h = \tfrac{1}{2}z \ln\left(1 + \frac{1}{\bar{v}}\right) = \frac{\hbar\omega}{2k\theta}z \tag{9.29}$$

$$\approx S \tag{9.29a}$$

and

$$\nu_h = \nu_{u0} + (\hbar\omega^2/2k\theta)z \tag{9.30}$$

$$\approx \nu_{u0} + S\omega. \tag{9.30a}$$

The peak value of G_h is

$$G_h(\max) = \frac{1}{(2\pi z)^{\frac{1}{2}}} \exp\left\{z\left(1 + \frac{\hbar^2\omega^2}{8k^2\theta^2}\right)\right\}. \tag{9.31}$$

Hence

$$G_h = G_h(\max) \exp\left\{-\frac{1}{2}\left(\frac{p}{\sqrt{z}} - \frac{\hbar\omega}{2k\theta}\sqrt{z}\right)^2\right\} \tag{9.32}$$

$$= G_h(\max) \exp\left\{-\frac{1}{2z\omega^2}(\nu - \nu_h)^2\right\}. \tag{9.32a}$$

Thus at high temperatures the absorption should be Gaussian provided (9.27a) and (9.27b) apply.

TABLE III.

θ °K	$\bar{v}$	$\frac{4k\theta}{\hbar\omega}$	$\frac{[\bar{v}(\bar{v}+1)]^{\frac{1}{2}}}{\theta}$	z for $S=30$
5	3.11×10^{-13}	0.139	1.11×10^{-7}	3.34×10^{-5}
10	5.58×10^{-7}	0.278	7.47×10^{-5}	4.48×10^{-2}
20	7.47×10^{-4}	0.555	1.36×10^{-3}	1.63
30	8.30×10^{-3}	0.834	3.04×10^{-3}	5.47
50	5.95×10^{-2}	1.39	5.02×10^{-3}	15.1
75	0.172	2.08	5.96×10^{-3}	26.8
100	0.311	2.78	6.40×10^{-3}	38.4
200	0.949	5.56	6.80×10^{-3}	81.6
400	2.31	11.1	6.91×10^{-3}	165.
600	3.69	16.7	6.93×10^{-3}	250.
800	5.07	22.2	6.96×10^{-3}	334.

If the ϵ_j's do not approach zero, p will be greater than unity, and criterion (9.27b) will hold, except in the region where $p=0$, i.e., $\nu = \nu_{u0}$. Since $p_h > 0$, this occurs to the red of the peak. In the broad bands which arise in polar solids, this will not be a real limitation on the applicability of (9.32a).

To test criterion (9.27a) at the peak, it is combined with (9.29) to give

$$2z/p_h = 4k\theta/\hbar\omega > 1. \tag{9.33}$$

In Table III some typical values of $\bar{v}$, z, $(4k\theta/\hbar\omega)$ and $(1/\theta)[\bar{v}(\bar{v}+1)]^{\frac{1}{2}}$ are given. The phonon frequency selected was $2\pi(3\times10^{12})$ sec^{-1} (M5). The table indicates that (9.33) applies above room temperature. Since $p < p_h$ for the red side of the band, Eq. (9.23a) should hold there.

If we multiply G_h by $(1/\omega)\exp\{-S(2\bar{v}+1)\}$ one sees that $\int G_n d\nu = 1$, provided one neglects small quantities. In the next section we show that this relation holds exactly.

Deviation from a Gaussian form should be looked for on the violet side of the peak where $p > p_h$; here (9.27b) could break down. This treatment does not suggest that a band should be Gaussian for all ν, although at medium temperatures one would suspect that the red side would approach such a shape. At some temperatures the following conditions may occur: $(2z/p) > 1$ for the red side (Gaussian), and $(z^2/4p) < 1$ on the violet side (Pekarian).

We may obtain an expression for H at high temperatures. Equations (7.7) and (9.32a) show that $2\hbar^2\omega^2 z = 1/a^2$ and

$$H = (5.545)^{\frac{1}{2}}\hbar\omega z^{\frac{1}{2}}$$

$$= (5.545)^{\frac{1}{2}}S^{\frac{1}{2}}\hbar\omega \operatorname{csch}^{\frac{1}{2}}\tfrac{1}{2}\beta. \tag{9.34}$$

For high θ

$$H = (11.090Sk\hbar\omega)^{\frac{1}{2}}\theta^{\frac{1}{2}}. \tag{9.34a}$$

Equation (9.30) predicts that the peak of the absorption band is temperature dependent. The dependence is given by $\nu_{u0}(\theta)$ and z/θ. In Table III the dependence of $[\bar{v}(\bar{v}+1)]^{\frac{1}{2}}/\theta$ on θ is given. This quantity is proportional to z/θ of (9.30). Except at low temperatures, where our approximations break down, this term is insensitive to variations of θ.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

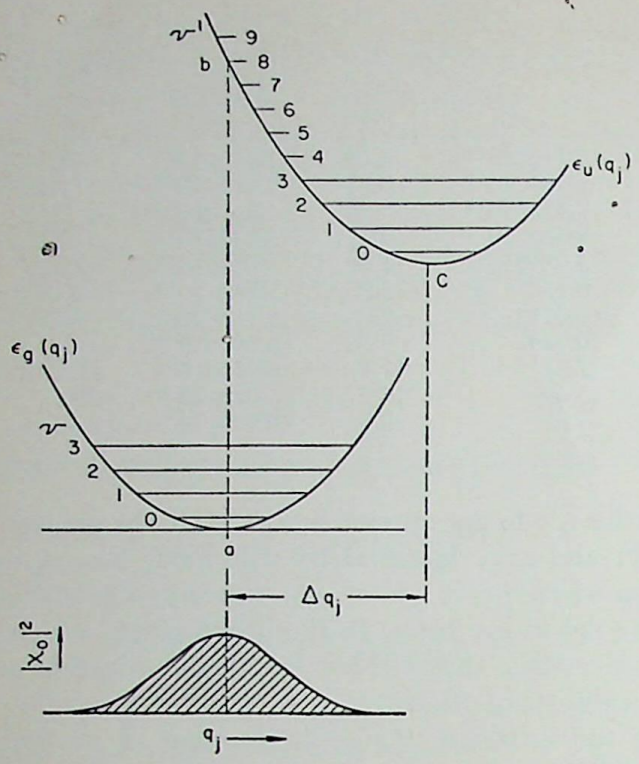


FIG. 6. Diagram illustrating the broadening effect. The upper and lower configurational coordinates are shown. The distribution probability at 0°K is given in the lower left corner. Point "a" is at $\epsilon_g(\mathbf{R}_g)$, while "c" is at $\epsilon_u(\mathbf{R}_u)$.

The dependence of ν_h and H on temperature given by (9.30) and (9.34) is due to the approximations used in this section. These relations require correction, but the high-temperature prediction is valid and (9.34a) holds.

The physical interpretation of the above equations is indicated in Fig. 6. Although the development considers more than one mode, this description is limited to the CC model. This is done for convenience only, as the arguments can be generalized to any number of effective modes. Since the ϵ_{jj}'s are zero, the frequencies are not affected by the transition. In Fig. 6 $\epsilon_g(q)$ and $\epsilon_u(q)$ are plotted against q. The horizontal lines are various values of $E=\epsilon_g+\hbar\omega(v+\frac{1}{2})$ and $E'=\epsilon_u+\hbar\omega(v'+\frac{1}{2})$. The relative probability of finding a particular value of q, when the electron is in the ground state, is shown by the shaded area on the lower left. The probable distribution, $\rho(q)$, at absolute zero is $\chi^2(q)$ for $v=0$. At any finite temperature it is

$$\rho(q)=\sum_v p(v)\chi^2(q)=\frac{1}{a\sqrt{\pi}}\exp[-q^2/a^2], \qquad (9.35)$$

where

$$a=\left(\frac{\hbar}{\omega}\right)^{\frac{1}{2}}\coth^{\frac{1}{2}}\tfrac{1}{2}\beta. \qquad (9.35a)$$

One obtains (9.35) from (8.14a) and (9.4). At high temperatures a^2 has the value $2k\theta/\omega^2$, and (9.35) has a form which one can obtain by Boltzmann's statistics if the total energy of the system is assumed to be $(1/2)\omega^2q^2$. The actual energy of the system, however, is not $(1/2)\omega^2q^2$, since (9.35) is the sum over many states unless $\theta=0$°K.

The first question is: what is S? It is simply the ratio of the difference in potential energy at 0 (b) and at Δq (c) to the energy of a phonon, $\hbar\omega$ (Fig. 2). This follows from (3.24) and (9.8a). The maximum of the absorption occurs for transitions from $v=0$ to $v'=S$, as is seen with the use of (9.22) and (9.29a). This is the vertical transition if we ignore the zero-point energy. One would expect this to be the peak because of the shape of $\rho(q)$. So far complex analysis has not led to anything very surprising.

The bandshape at low temperature, however, is non-Gaussian; indeed it is nonsymmetric. This means that $|\{\chi_{S+1}'|\chi_0\}|^2$ does not equal $|\{\chi_{S-1}'|\chi_0\}|^2$. Here 0 and $S\pm1$ refer to the quantum states of χ and χ'. This lack of symmetry leads to the skewness in absorption curves at low temperatures. This has created a serious problem in the experimental analysis (K2a, p. 119). We stress that this lack of symmetry stems from the properties of the χ's and not $\epsilon_u(q)-\epsilon_g(q)$, which for this problem has the simplest form possible, namely

$$\hbar\nu=\hbar\nu_{ug}+\tfrac{1}{2}\omega^2(\Delta q)^2-\omega^2(\Delta q)q. \qquad (9.36)$$

ν is a *linear* function of q.

At high temperatures the spread on the red side occurs because of transitions where $v_t'-v_t$ is smaller than S. The violet side is affected by transitions from the ground state where v_t' is greater than S. The theory suggests that these transitions tend to make the curve more symmetrical and in the extreme case, Gaussian.

The Williams' approximation suggests that the transition probability at frequency ν is obtained from (9.35) and (9.36). First, (9.36) is solved for q; then the absorption is obtained from $\rho(q)$. Since $\rho(q)$, however, is composed of several states (unless $\theta=0$°K), this approach is inconsistent, since $\epsilon_u(q)-\epsilon_g(q)$ does not equal $E'-E$. Further, the matrix elements are somewhat asymmetrical.

The question of the asymmetry of the absorption band is an extremely fundamental one. The Williams' approximation happens to give the correct temperature dependence of H (Sec. 10). Equation (9.34) might also be a good approximation. Actually, it is a poor one. It is the asymmetry of the curves, i.e., Eq. (9.21), which gives us the necessary insight to understand what is really happening.

In conclusion, one should mention that Krivoglaz and Pekar (K5) report a shape where the term const$(\nu-\nu_h)^3$ is added to the expression in the curly brackets of Eq. (9.32a). This term will create a skewness when $(\nu-\nu_h)$ is not too small. As ν approaches ν_h, the curve becomes symmetric and almost Gaussian. This fits the experimental data better than (9.32a), since there is another adjustable parameter. The experimental data seems to resemble a double Gaussian curve more nearly, however. There is no evidence, however, to believe that a discontinuity in the absorption occurs at $\nu=\nu_h$.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

Arguments by Kubo and Toyozawa (K6) indicate that the simple model employed here gives a symmetrical shape at high temperatures. This same conclusion can be obtained from the higher moments calculated by O'Rourke. The conclusion of Sec. 9(b) regarding the symmetry of the shape is not due to (9.25). Experimental data suggests (K4) that the results may not be completely valid. However, the simplifying assumptions give some useful information regarding the low-temperature shape and the transition probabilities.

10. METHOD OF MOMENTS

The method of moments was developed by Lax (L1) and extended by Meyer (M8). We now combine the development of Part I with this powerful approach.

Returning to Eq. (8.13a), we first sum over v_t' and then take a weighted average over v_t using (8.14); thus,

$$\alpha(\nu)=\frac{K}{2\pi}\,\mathrm{Av}\int_{-\infty}^{\infty}\sum_{v'}dt\{\chi|f^{\frac{1}{2}}|\chi'\}\{\chi'|f^{\frac{1}{2}}|\chi\}$$

$$\times\exp\left\{i\left[(E'-E)\frac{1}{\hbar}-\nu\right]t\right\}, \quad (10.1)$$

$$K=2\pi^2be^2N/cm. \quad (10.1a)$$

In this approach $\alpha(\nu)$ is used. To obtain (10.1) we have made use of (9.1). Av implies a weighted average over the ground vibrational states. By expanding $\exp\{iE't/\hbar\}$ in series, one may show that

$$\chi'\exp\{iE't/\hbar\}=e^{iH_ut/\hbar}\chi'. \quad (10.2)$$

Now (10.1) takes the form

$$\alpha(\nu)=\frac{K}{2\pi}\,\mathrm{Av}\int dt\sum_{v'}$$

$$\times\{\chi|f^{\frac{1}{2}}e^{iH_ut/\hbar}|\chi'\}\{\chi'|f^{\frac{1}{2}}e^{-iH_gt/\hbar}|\chi\}e^{-i\nu t}. \quad (10.3)$$

We note that E' depends on v_t'; H_u, however, does not. Therefore, we take it out of the sum. Equation (10.3) is a product of the three integrals: one over t; the other over the nuclear coordinates involved in the first { }; while the third arises because of the second { }. Since the χ's form a complete set in q space, we use the following property of eigenfunctions (H3, p. 66)

$$\sum_{v_j'}\chi_j'(q_j)\chi_j'(q_j')=\delta(q_j-q_j') \quad (10.4)$$

and write

$$\alpha(\nu)=\frac{K}{2\pi}\int dt\,\mathrm{Av}\{\chi|f^{\frac{1}{2}}e^{iH_ut/\hbar}f^{\frac{1}{2}}e^{-iH_gt/\hbar}|\chi\}e^{-i\nu t}. \quad (10.5)$$

Now the Condon approximation is used to remove f from the matrix element, giving

$$\alpha(\nu)=\frac{Kf}{2\pi}\,\mathrm{Av}\int dt\{\chi|e^{i(H_u-H_g)t/\hbar}|\chi\}e^{-i\nu t}. \quad (10.6)$$

The quantum mechanics is contained in the order of the operators; thus one cannot remove $\exp\{i(H_u-H_g)t/\hbar\}$ from the matrix (Appendix V).

By assuming that f is frequency independent, we see that α is the Fourier transform of the function $Kf\,\mathrm{Av}\{\chi|\exp[i(H_u-H_g)t/\hbar]|\chi\}$ and it follows that (M7, p. 252)

$$Kf\,\mathrm{Av}\{\chi|e^{i(H_u-H_g)t/\hbar}|\chi\}=\int\alpha(\nu)e^{i\nu t}d\nu. \quad (10.7)$$

Setting $t=0$, we find

$$\int\alpha(\nu)d\nu=Kf. \quad (10.8)$$

Therefore, it follows that $G_n(\nu)$ of (8.18a) is normalized. The thermal vibration spreads the absorption band but does not affect the area. Equation (10.8) is based on the approximation that f is *independent* of frequency. Had we assumed that $\langle\varphi'|\,\mathbf{r}\,|\varphi\rangle$ is frequency dependent, the zero moment would be temperature sensitive. Using (7.1), it follows that

$$M_0=\hbar Kf. \quad (10.9)$$

Further,

$$\left\{(\hbar/i)(\partial/\partial t)\int\alpha(\nu)e^{i\nu t}d(\hbar\nu)\right\}_{t=0}=\int\alpha(\nu)\hbar\nu d(\hbar\nu)$$

$$=(\hbar^2/i)Kf\,\mathrm{Av}[(\partial/\partial t)\{\chi|e^{i(H_u-H_g)t/\hbar}|\chi\}]_{t=0} \quad (10.10)$$

and

$$M_1=\hbar Kf\,\mathrm{Av}\{\chi|H_u-H_g|\chi\}. \quad (10.10a)$$

Now,

$$\left\{-\hbar^2\frac{\partial^2}{\partial t^2}\int\alpha(\nu)e^{i\nu t}d(\hbar\nu)\right\}_{t=0}=\int(\hbar\nu)^2\alpha(\nu)d(\hbar\nu) \quad (10.11)$$

and

$$M_2=\hbar Kf\,\mathrm{Av}\{\chi|(H_u-H_g)^2|\chi\}. \quad (10.14)$$

By similar methods we may evaluate the higher moments. For these situations, one may not write

$$\exp\{(i/\hbar)(H_u-H_g)t\}$$

for

$$\exp\{(i/\hbar)H_ut\}\exp\{(-i/\hbar)H_gt\}$$

as explained in Appendix V.

To evaluate (10.10) and (10.12), expressions for the operator H_u-H_g as well as $\mathrm{Av}\{\chi|q_j^n|\chi\}$ (n is an integer) are required. The matrix elements can be

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

calculated from the properties of the χ's and are (I, p. 173–181)

$$\mathrm{Av}\{\chi|q_j|\chi\}=0 \tag{10.13a}$$

$$\mathrm{Av}\{\chi|q_j^2|\chi\}=[\hbar/2\omega_j(g)]\coth\tfrac{1}{2}\beta_j \tag{10.13b}$$

$$\mathrm{Av}\{\chi|q_j^3|\chi\}=0 \tag{10.13c}$$

$$\mathrm{Av}\{\chi|q_j^4|\chi\}=3[\hbar/2\omega_j(g)]^2\coth^2\tfrac{1}{2}\beta_j. \tag{10.13d}$$

β_j is defined by (8.14b) where the ground state ω_j is used.

An expression for (H_u-H_g) is obtained from (8.1), (8.2) and the expansion technique employed in Sec. 3. It now follows that

$$H_u-H_g=\Delta\epsilon(\theta)+\textstyle\sum_j \epsilon_j q_j+\frac{1}{2}\sum_j \epsilon_{jj}q_j^2 +\frac{1}{2}\sum_{jk}{}' \epsilon_{jk}q_jq_k \tag{10.14}$$

$$\Delta\epsilon(\theta)=\epsilon_u(\mathbf{R}_g)-\epsilon_g(\mathbf{R}_g)=\hbar\nu_{ug}+\tfrac{1}{2}\textstyle\sum_j \epsilon_j^2/\omega_j^2(u). \tag{10.14a}$$

To obtain (10.14) and (10.14a) use has been made of (3.27). When comparing Secs. 9 and 10, it should be remembered that $\Delta\epsilon$ does *not* equal $\hbar\nu_{ug}$. The prime on the sum means that $j\neq k$. The temperature dependence of $\Delta\epsilon$ is due to the expansion of the crystal. We assume that the ϵ_j's and ϵ_{jk}'s are independent of θ. The derivatives are evaluated at the equilibrium position of the ground state, and the q_j's are the normal modes for this state.

The evaluation of M_1 and M_2 is now straightforward. Using (10.9) and (10.10)

$$\bar{\epsilon}=(M_1/M_0) =\Delta\epsilon(\theta)+\tfrac{1}{4}\textstyle\sum_j \epsilon_{jj}[\hbar/\omega_j(g)]\coth\tfrac{1}{2}\beta_j. \tag{10.15}$$

To evaluate M_2 we square (10.14), i.e.,

$$\begin{aligned}(H_u-H_g)^2=&\Delta\epsilon^2(\theta)+\textstyle\sum_j \epsilon_j^2q_j^2+\sum_{jk}{}' \epsilon_j\epsilon_k q_jq_k\\ &+\tfrac{1}{4}\textstyle\sum_j \epsilon_{jj}^2q_j^4+\tfrac{1}{4}\sum_{jk}{}' \epsilon_{jj}\epsilon_{kk}q_j^2q_k^2\\ &+2\Delta\epsilon(\theta)\textstyle\sum_j \epsilon_jq_j+\Delta\epsilon(\theta)\sum_j\epsilon_{jj}q_j^2\\ &+\Delta\epsilon(\theta)\textstyle\sum_{jk}{}' \epsilon_{jk}q_jq_k\\ &+\sum_{jk}\epsilon_j\epsilon_{kk}q_jq_k^2,\end{aligned} \tag{10.16}$$

where again terms involving $\epsilon_j\epsilon_{kl}$ and $\epsilon_{ij}\epsilon_k\epsilon_l$ $(k\neq l)$ have been omitted. Using (10.12) and (10.13) results in

$$\begin{aligned}(M_2/M_0)=&\Delta\epsilon^2(\theta)+\textstyle\sum_j[\hbar/2\omega_j(g)]\epsilon_j^2\coth\tfrac{1}{2}\beta_j\\ &+\tfrac{1}{8}\textstyle\sum_j\epsilon_{jj}^2[\hbar^2/\omega_j^2(g)]\coth^2\tfrac{1}{2}\beta_j\\ &+\tfrac{1}{16}\textstyle\sum_{jk}\epsilon_{jj}\epsilon_{kk}[\hbar^2/\omega_j(g)\omega_k(g)]\coth\tfrac{1}{2}\beta_j\coth\tfrac{1}{2}\beta_k\\ &+\tfrac{1}{2}\Delta\epsilon(\theta)\textstyle\sum_j\epsilon_{jj}[\hbar/\omega_j(g)]\coth\tfrac{1}{2}\beta_j.\end{aligned} \tag{10.17}$$

A combination of (7.3), (10.15), and (10.17) gives the important equation

$$\begin{aligned}m^2=&\textstyle\sum_j\epsilon_j^2[\hbar/2\omega_j(g)]\coth\tfrac{1}{2}\beta_j\\ &+\tfrac{1}{8}\textstyle\sum_j\epsilon_{jj}^2[\hbar^2/\omega_j^2(g)]\coth^2\tfrac{1}{2}\beta_j.\end{aligned} \tag{10.18}$$

We have made only two assumptions—that f is not a function of ν or q, and that the adiabatic approximation can be used to terms of ϵ_{ij}, ϵ_{kl}, and $\epsilon_j\epsilon_{kl}$, where $k\neq l$.§§§ These equations are extremely general. The possibility of calculating the q's and ϵ's about a point imperfection with any degree of accuracy is small. Hence, the equations as they stand are of little use. We must make further assumptions to obtain relations between $\bar{\epsilon}$, m, and θ which involve parameters to be evaluated experimentally. The generality of the equations assures one that terms will not be omitted. Thus, we may apply the equations to the Huang-Rhys problem (the Fröhlich approximation) and compare our results with those of O'Rourke and Meyer. On the other hand, we may make the approximation used by Klick.

(a) CC Model

Here, there is only one q whose ground state has the angular frequency $\omega(g)$; hence, from Sec. 3,

$$X=q/M^{\frac{1}{2}}, \tag{10.19}$$

where M is an effective mass of the order of an ionic mass. For this model

$$\begin{aligned}\Delta\epsilon(q)=&\hbar\nu_{ug}+\tfrac{1}{2}\omega^2(u)(q-\Delta q)^2-\tfrac{1}{2}\omega^2(g)q^2\\ =&\hbar\nu_{ug}+\tfrac{1}{2}\omega^2(u)(\Delta q)^2-\omega^2(u)(\Delta q)q\\ &+\tfrac{1}{2}[\omega^2(u)-\omega^2(g)]q^2.\end{aligned} \tag{10.20}$$

Using the definition of (3.18),

$$\epsilon_j=-\omega^2(u)\Delta q=-\omega^2(u)M^{\frac{1}{2}}\Delta X \tag{10.21a}$$

and

$$\epsilon_{jj}=\omega^2(u)-\omega^2(g)\approx 2\omega(g)\Delta\omega. \tag{10.21b}$$

Equations (10.15) and (10.18) now have the form

$$\bar{\epsilon}=\Delta\epsilon(\theta)+\tfrac{1}{2}\hbar\Delta\omega\coth\tfrac{1}{2}\beta \tag{10.22a}$$

and

$$m^2=\left\{\frac{1}{2}\frac{\omega^2(u)}{\hbar\omega(g)}M(\Delta X)^2\right\}\hbar^2\omega^2(u)\coth\tfrac{1}{2}\beta \times+\tfrac{1}{2}\hbar^2(\Delta\omega)^2\coth^2\tfrac{1}{2}\beta. \tag{10.22b}$$

A generalized S is defined by the use of (10.18) and (10.22b),

$$S=\frac{1}{2\hbar}\sum_j\epsilon_j^2\frac{1}{\omega_j(g)\omega_j^2(u)}=\tfrac{1}{2}\sum_j\frac{\omega_j^2(u)}{\hbar\omega_j(g)}(Q_j-q_j)^2. \tag{10.23}$$

The last step uses (3.24). S equals the expression in the curly brackets of (10.22b). This definition agrees with (9.8a) and is equivalent to the one used by Meyer. An exact comparison is somewhat involved since Meyer takes a thermal average over modes which do not include V_I and V_e of (5.2). The S for absorption and emission need not be the same. Both $\omega(g)$ and $\omega(u)$ enter into the definition.

§§§ We also assume that the modes are nondegenerate (Sec. 3).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

The second term in (10.22b) is new; it arises from a superior approximation.‖‖‖ To estimate its effect we set $\theta=0°\text{K}$ and $\Delta\omega\approx\omega$. Now

$$m_0=\hbar\omega[S+(1/2)]^{\frac{1}{2}}$$
$$=\hbar\omega S^{\frac{1}{2}}[1+(1/4S)]. \quad (10.22c)$$

If S is of the order of 30, the correction term is of the order of 0.01, which is at the experimental limit of accuracy. If S were smaller than 30, and if $\Delta\omega$ is large, one might be able to detect the second term in (10.22b). This would be done from the temperature behavior of m.

Equation (10.22a) has a great deal less information than (10.22b), since no reliable information regarding $\Delta\epsilon(\theta)$ exists. As stated, $\Delta\epsilon$ may change because of the lattice expansion. ϵ_m for the F center seems to be a function of the interionic distance, d. From empirical data, one may obtain the following relation, $\epsilon_m(d)=\text{const}\ d^{-1.84}$ (Mollwo (M10), Ivey (I1)). Combining this with density measurements at various temperatures, one may calculate $\Delta\epsilon(\theta)$. Using Henglein's (H3a) expansion data for KCl, one obtains

$$\epsilon_m(193°\text{K})-\epsilon_m(273°\text{K})=0.011\ \text{ev}$$
$$\epsilon_m(80°\text{K})-\epsilon_m(273°\text{K})=0.025\ \text{ev},$$

where $\bar{\epsilon}$ has been replaced by ϵ_m. This change of ϵ_m is too small to explain the observed temperature variation (0.040 ev and 0.070 ev), which is an indication that $\Delta\omega\neq0$ for this particular absorption band. One might attempt to combine the Ivey relation with (10.22a).

On the other hand, we may assume that $\Delta\epsilon(\theta)$ equals a constant. This can be tested by plotting ϵ_m against $\coth\beta/2$; β having been evaluated empirically from (10.22b). If the data give the equation $A+B\coth\beta/2$, one may evaluate $\Delta\epsilon$ and $\Delta\omega$. There is a danger in this procedure since $\Delta\epsilon(\theta)$ may also vary as $A+B\coth\beta/2$ over a limited range of temperatures [for small β, $\coth\beta/2\approx(2/\beta)$]. In this situation the coefficient of $\coth\beta/2$ would not equal $\frac{1}{2}\hbar\Delta\omega$. As the range of θ increases, the possibility of accidental correlation decreases.

Using recent data on the F center in KCl, ω has been evaluated with the use of (10.22b) (without the second term). From (10.22a) ω, $\bar{\epsilon}$, and $\hbar\Delta\omega$ were determined. One again concludes that the correction to (10.22b) is very small.

Klick and Schulman (K2a, p. 110) have used the Williams' approximation to obtain an expression for H. It is identical to the first term of (10.22b), provided H and m are proportional. This does not imply that the derivation given here is identical to the one obtained previously. The latter method does not give the second term of Eq. (10.22b), nor the Pekarian shape (at low temperature). The latter is readily observable and requires understanding. If one uses Eq. (9.34), H depends on the $\text{csch}\beta/2$ instead of $\coth\beta/2$, which indicates that the relation given by William and Hebb (W3) and by Lax (L2) is fortuitous.

‖‖‖ A term of this nature appears in the formal treatment of Kubo and Toyozawa (K6).

(b) Huang-Rhys Problem

In this case we assume that all the modes have a single frequency given by the LST relation; then

$$\bar{\epsilon}=\Delta\epsilon(\theta)+\tfrac{1}{2}\textstyle\sum_j\hbar\Delta\omega_j\coth\frac{1}{2}\beta \quad (10.24a)$$

and

$$m^2=\hbar^2\omega^2(u)S\coth\tfrac{1}{2}\beta+\tfrac{1}{2}(\textstyle\sum_j\hbar^2\Delta\omega_j^2)\coth^2\tfrac{1}{2}\beta. \quad (10.24b)$$

Equation (10.24a) was first obtained by O'Rourke, using the approximation that $\Delta\omega$ is *very* small. Later Meyer obtained this equation by an alternate method. These authors did not note the thermal dependence of $\Delta\epsilon(\theta)$. The first term of Eq. (10.24b) has been given before, while the second term is new, see however (K6). It may be detected when careful measurements are made on bands with small S's.

(c) Double-Frequency Model

We may apply (10.15) and (10.18) to more complex situations. An alternate model is one with two effective ground frequencies ω_1 and ω_2. For this situation

$$\bar{\epsilon}=\Delta\epsilon(\theta)+\frac{\hbar}{4\omega_1(g)}\{\textstyle\sum_j\epsilon_{jj}\}\coth\frac{1}{2}\beta_1$$
$$+\frac{\hbar}{4\omega_2(g)}\{\textstyle\sum_i\epsilon_{ii}\}\coth\frac{1}{2}\beta_2 \quad (10.25a)$$

and

$$m^2=\hbar^2\omega_1^2(u)S_1\coth\tfrac{1}{2}\beta_1+\hbar^2\omega_2^2(u)S_2\coth\tfrac{1}{2}\beta_2. \quad (10.25b)$$

Here

$$S_1=\tfrac{1}{2}\textstyle\sum_j[1/\hbar\omega_j(g)\omega_j^2(u)]\epsilon_j^2 \quad (10.26a)$$

$$S_2=\tfrac{1}{2}\textstyle\sum_i[1/\hbar\omega_i(g)\omega_i^2(u)]\epsilon_i^2 \quad (10.26b)$$

$$\beta_1=(\hbar/k\theta)\omega_1(g)\quad\text{and}\quad\beta_2=(\hbar/k\theta)\omega_2(g). \quad (10.26c)$$

j is summed over the modes associated with the frequency ω_1, and i is over the ω_2 modes. The second term in (10.18) is omitted.

We again use an effective frequency and write

$$S_1=[1/2\hbar\omega_1(g)\omega_1^2(u)]\textstyle\sum_j\epsilon_j^2. \quad (10.27)$$

Since the data are always limited to a range in temperature, this may be less restrictive than (9.7) which must hold for every value of t. If (10.27) is to hold for every value of β_j, there cannot be a spread in the values of the ω's.

We may assume that S_1 is related to the local modes, and S_2 is associated with the longitudinal optical modes. β_1 will be roughly half β_2. The question is: can one detect the difference between (10.25a and b) and (10.22a and b) or (10.24a and b)? For the answer to this question, we must consider the various methods employed to analyze the data.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

11. ANALYSIS OF EXPERIMENTAL DATA

In the last two sections, we derived several equations which can be compared with experimental data. Various attempts have been made to analyze the data with radically different results. Some of the difficulties stem from the data itself, yet this is not the major problem. While at low temperatures (Table I) there is considerable disagreement, this is not the case at room temperature (Table IV). The agreement between Mollwo's (M10) and the most recent data (M5) is good. Russell and Klick's (R1) value of H seems to be too large at all temperatures.

Russell and Klick conclude that (10.22b) (without the second term) applies to the F center, provided the angular frequency in KCl is 1.6×10^{13} sec^{-1}. On the other hand, Pekar (P2 and 4) and Meyer (M8 and 9) conclude that the same equation holds, provided a value of 3.95×10^{13} sec^{-1} is used. It was obtained from the LST relation. Pekar and Meyer use Mollwo's data. The reason for the difference stems from the method of treating the data. One may use Mollwo's data with an alternate method of analysis and obtain 2×10^{13} sec^{-1} for ω.

Pekar's analysis is most remarkable (P4, p. 129 ff and p. 145 ff). He had only one adjustable constant, the

TABLE IV. F center in KCl at room temperature.

Experimenter	(M10)	(R1)	(M5)
ϵ_m ev	2.20	2.2	2.225
H ev	0.35	0.39	0.35

effective mass of the electron. It was evaluated from the experimental ϵ_m at room temperature; hence, his results should apply to the most recent data. Using this value, Pekar calculated S by employing (9.8a) and (3.18a). For the relation between the q's and the X's he used the very long wave approximation (B, p. 213). His value of S for KCl is 23.8. Using this value, the LST relation (10.24b), and (7.7b), we compute the values in Table V. Equation (7.7b) was used to relate m and H; this gives

$$H^2=8\ln 2(\hbar\omega)^2\coth\tfrac{1}{2}(\hbar\omega/k\theta). \tag{11.1}$$

The calculation of H at 300°K does not agree with the value, 0.40, reported in Pekar's paper of 1950 since it was based on Eq. (9.34), not (10.24b). Strangely enough, in 1953 Krivoglaz and Pekar (K5) derived (11.1). However the corrected numerical values of H were not reported in Pekar's subsequent review paper (P3), although (11.1) was quoted. The correct expression for H at 0°K seems to have been first given by Pekar in 1950 (P1). While the results of the table are not satisfactory, they are really quite remarkable, since there is only *one* adjustable parameter.

A method of analysis which the author believes is quite useful is now discussed. If one has to use Eqs. (10.15) and (10.18) with various values of β_j, the problem becomes hopeless since the number of adjustable parameters is too large. The opposite point of view, which is very naive, is that one exists and that Eqs. (10.22a) and (10.22b) or (10.24a) and (10.24b) apply. If we further assume that $\Delta\epsilon$ is independent of temperature, there are only four adjustable parameters and the equations of interest take the forms¶¶¶

$$\bar{\epsilon}=\Delta\epsilon+B\coth\tfrac{1}{2}(\hbar\omega/k\theta) \tag{11.2}$$

$$m^2=\hbar^2\omega^2S\coth\tfrac{1}{2}(\hbar\omega/k\theta)+\tfrac{1}{2}B^2\coth^2\tfrac{1}{2}(\hbar\omega/k\theta). \tag{11.3}$$

Next we may assume that B^2 is very small compared to $\hbar^2\omega^2S$ and hence can be omitted from (11.3). This is true for the F center in KCl. To obtain m from an arbitrary curve is a laborious process involving a great deal of numerical work. In some cases, one may be able to replace m by H [Eqs. (7.7b) and (7.12b)], hence in the simplest situation we wish to test critically the relation

$$H^2=H^2(0)\coth\tfrac{1}{2}(\hbar\omega/k\theta), \tag{11.4}$$

where $H(0)$ is the half-width at 0°K. It can readily be obtained from measurements at low temperatures (approx 4°K). The validity of (11.4) is based on the assumption that the ratio of m to H is temperature independent. This has to be tested by numerical integration.

Equation (11.4) suggests plotting $\coth^{-1}[H/H(0)]^2$ against $1/\theta$. We have the desired test if $1/\theta$ can be varied over a large range of values. A straight line going through the origin is a clear indication of the validity of (11.4) and the development in Sec. 10. Further, it shows that there is one effective mode, so the theory developed in Sec. 9 can be used. Such a plot is a more critical test of the theory than the previous methods of analysis, since one deals with straight lines rather than a complex curve. If one combines this technique of analysis with Mollwo's data, one obtains the 2×10^{13} sec^{-1} value mentioned above.

Three alternate methods have been suggested.

(1) Klick (K2) plots H against $\theta^{\frac{1}{2}}$. At high temperatures, a straight line is obtained whose slope is proportional to $S^{\frac{1}{2}}\omega^{\frac{1}{2}}$. At moderately low temperatures, a departure from a straight line is obtained and the analysis requires the determination of this departure. At extremely low temperatures (about 25°K), one gets a straight horizontal line. While Klick's approach is useful, one cannot determine the validity of (11.4) in a

TABLE V. Comparison of Pekar's parameters with experiment.

θ	H (cal)	H (exp)
300°K	0.43	0.35
0°K	0.30	0.16

¶¶¶ We assume here that $(\Sigma_j\hbar\Delta\omega_j)^2=\Sigma_j(\hbar\Delta\omega_j)^2$ which, of course, need not be true except in the CC model. The B from (11.2) will however give us the order of magnitude of B^2 in (11.3).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

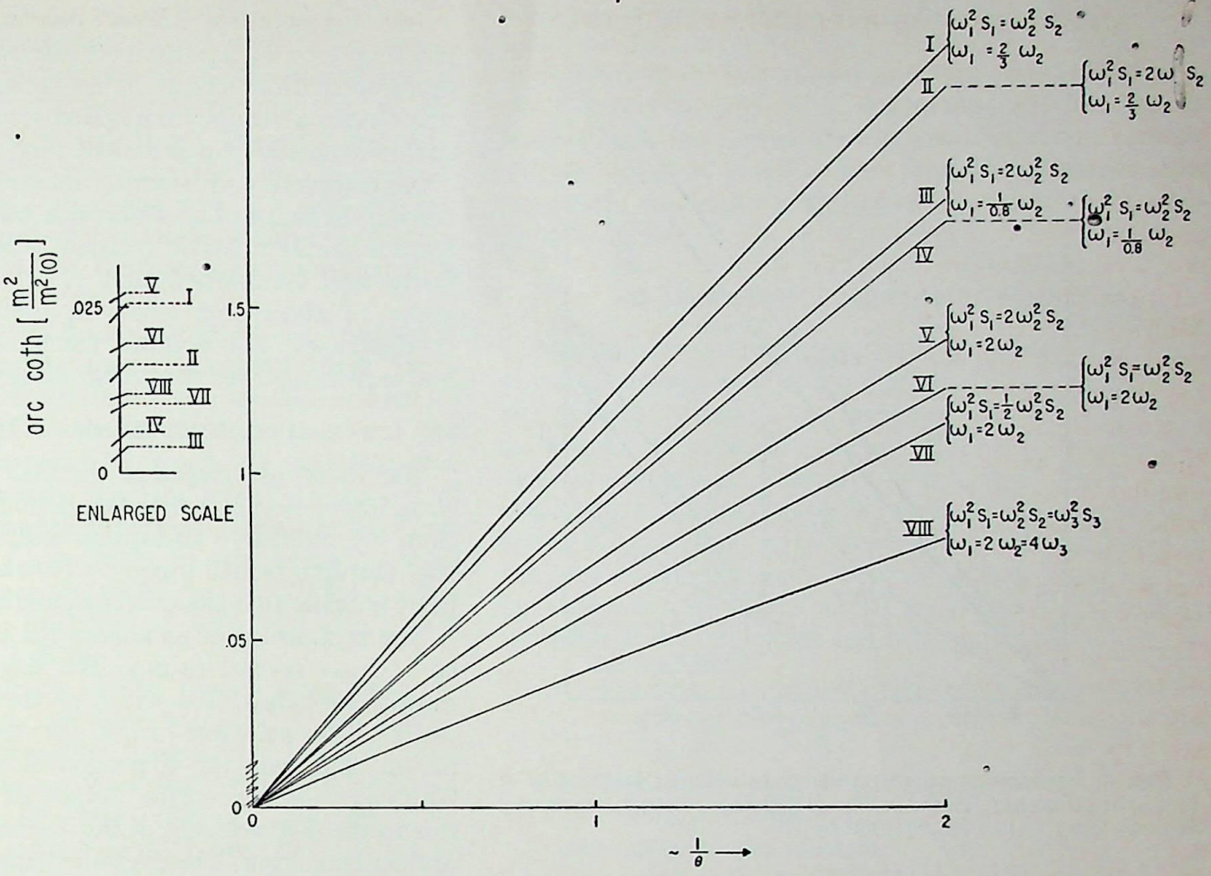


FIG. 7. Broadening caused by more than one effective frequency. Note the curve at $1/\theta=0$. In some cases m can be replaced by H. Analysis of more than one ω.

critical manner since the points are not supposed to lie on a straight line.

(2) Pekar (P3) has plotted H against θ. In this case one does not get a straight line except at extremely low temperatures. An analysis of experimental data by this method is difficult, and the author knows of no detailed study using this approach.

(3) Meyer (M8) has plotted $(H/\hbar\omega)^2$ against $\coth\frac{1}{2}(\hbar\omega/k\theta)$ using a known ω (from the LST relation). An erroneous ω can be detected only at low temperatures (under 100°K). If a great many points are available below 100°K, one can see that his technique does not give a straight line except for the correct value of ω. Thus, using recent data on the F center in KCl (M5) one may show that his ω is incorrect. Mollwo's data had too few points in the critical region, so that Meyer was unable to detect the deviation. Actually, Meyer was unable to confirm (11.4) and was forced to add a constant [Appendix V and Eq. (11.5)].

We are particularly interested in what happens when there is a breakdown in the approximation made in arriving at (11.4). In particular, how does Eq. (10.25b) look on a plot of $\coth^{-1}[H^2/H^2(0)]$ against $1/\theta$? This question is only meaningful provided the ratio of ω_1 to ω_2 is not near unity, and S_1 and S_2 are approximately equal. One cannot hope to detect the difference between Eqs. (10.25b) and (11.4) with the use of experimental data unless there is an appreciable contribution from both types of modes, and the frequencies are distinct. One may assume that both the local modes and the longitudinal optical modes interact with the trapped electron. The local modes should be in the gap between the acoustical and optical branches. In NaCl and KCl this gap is roughly at one-half the frequency of the longest longitudinal optical modes; hence the ratio of the ω's should be of the order of 2.

To understand the problem various $\coth^{-1}[H(\theta)/H(0)]^2$ against $1/\theta$ plots have been made using several ratios of S_1 to S_2 and several ratios of ω_1 to ω_2 (Fig. 7).**** In making these graphs, the values of S_1, S_2, ω_1, and ω_2 were assumed to be known. Then, $H^2(\theta)$ was calculated. From this H^2, $\coth^{-1}[H^2(\theta)/H^2(0)]$ was computed. Since the data refer only to a limited range of temperatures, the curvature shown in Fig. 7 will not be detected. A sure indication of the difference between (11.4) and (10.25b) is in the intercept, where $1/\theta=0$. If the line goes through the origin, we have a fair experimental proof that (11.4) applies, and that there is a *single effective* frequency. If not, one must attempt to express H by means of a more complete equation. Strangely enough, the data on the F center in KCl can be interpreted in this very simple manner (K4).

Similarly, in Fig. 8 plots have been made with the use of Meyer's equation, namely,

$$H^2=C+D\coth\tfrac{1}{2}(\hbar\omega/k\theta). \qquad (11.5)$$

This equation assumes that $f^{\frac{1}{2}}$ is a linear function of q

**** In Figs. 7 and 8 we assume that $m\approx H$.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

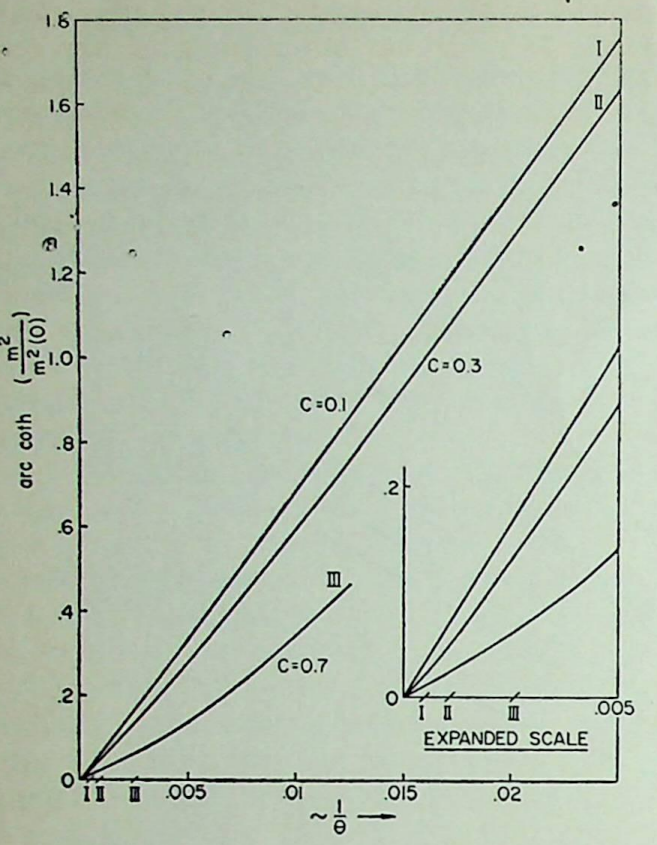


FIG. 8. Broadening effects when f^2 is a linear function of q. Note the curve at $1/\theta=0$ [in Eq. (11.5) D is set equal to unity]. In some cases m can be replaced by H.

(Appendix V). Again, the same conclusion is obtained; namely, the validity of (11.5) can be judged by the interception at $1/\theta=0$.

Use of this kind of plot requires measurements at both low and high temperatures. Measurements at extremely low temperatures cannot be used, since $H^2/H^2(0)$ approaches unity, and there is a large uncertainty in the $\coth^{-1}$ of $H^2/H^2(0)$ due to slight uncertainty in the evaluation of H^2 (one or two parts per thousand). This is one disadvantage of this plot. The other proposed methods are no better in this temperature region if (11.4) applies.

As suggested by Meyer, one may test (11.2) by plotting $\bar{\epsilon}$ or ϵ_m against $\coth\frac{1}{2}(\hbar\omega/k\theta)$, ω having been determined from (11.4). If $\Delta\epsilon$ is independent of θ one will obtain a straight line. Of course, $\Delta\epsilon$ could vary as const$+b\coth\beta/2$ for a limited range of θ.

No completely satisfactory method suggests itself for testing the shape of an absorption band, since we do not have a simple analytic relation between α and ϵ. It is fairly well established that (7.7) and (7.8) will not apply to absorption bands in solids, except in very special cases. For detailed comparison of the theory and experiment, see the forthcoming paper (K4).

12. CONCLUSIONS OF PART II

The theory developed in the first sections is applied to an impurity center with two bound states. Considerations are limited to absorption and, for a large part, to the simplest model (Secs. 9, 11, and 17). It has thus been possible to make exact calculations.

The broadening of an absorption band is on account of at least several causes. The band is composed of narrow subbands. The width of these arises from the uncertainty principle, from the center-center interaction, and from the dispersion of the normal modes. The envelope curve (the height of the subbands) is due to the displacement of the normal modes during an optical transition. The over-all shape arises from terms which appear in phonondynamics and are absent in electrodynamics. The broadening effects are illustrated in Fig. 9. The energy levels of the system are drawn on the left-hand side. At absolute zero before the transition, the system is in its ground state ($v_t=0$). Various transition probabilities exist for the excitation to levels in the excited state ($v_t'=0, 1, 2, \cdots$). The height of the vertical lines on the right are proportional to these transition probabilities (Eq. 9.21). The result is an asymmetric curve which resembles the F center in KCl (K4). We have plotted v_t' from right to left as is usually done. For emission one may exchange v_t and

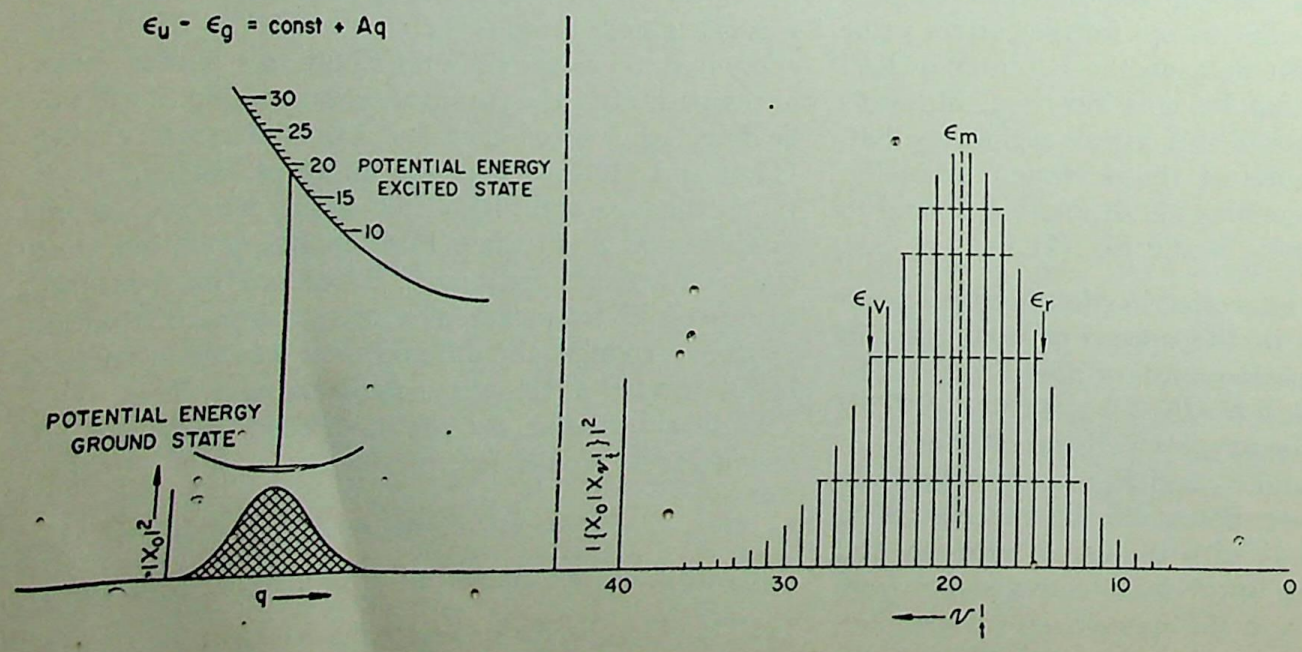


FIG. 9. Band shape at $\theta=0°$K for $S=20$. The vertical lines on the right are proportional to the transition probability. Note $\epsilon_v-\epsilon_m>\epsilon_m-\epsilon_r$ and the lack of symmetry for all values of α.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

v_t'. Here the emitted energy is $\Delta\epsilon(Q=0)-\hbar\omega(v_t-v_t')$, and the diagram should be "flipped over." The absorption curve shown on the right is probably not limited to the model on the left but only requires that all the important normal modes (whose ϵ_j do not equal zero) have an effective single frequency (Sec. 9). The generalization of Fig. 9 to any model with a single effective frequency is a result of the use of Mehler's formula. This is a result of the fundamental contribution of O'Rourke.

The shape presented is thus exact for the model considered at absolute zero. To extend these calculations to higher temperatures, various approximations must be made. The asymmetry disappears and the agreement with experiment is much less good. One may indeed show rigorously that the model on the left of Fig. 9 gives a symmetrical curve at high temperatures. This can be done most simply by calculating the third moment (01). Hence the model has to be expanded to predict the shape observed at high temperatures in the F center of KCl (K4).

As to the accuracy of the theory developed, the simple model (Sec. 9) does not predict the correct relation between E^e and E^a. For example, in the F center of KCl, $S=30$, $\hbar\omega=0.012$ ev, and E^a (at 0°K) $=2.3$ ev. From (4.2a), $E^e=E^a-2\hbar\omega S=1.6$ ev compared to the experimental value of 1.2 ev (V1). For this situation we may not let the frequency change be zero, and a more complex calculation is required (Sec. 10). For large values of S, we do not have to describe the potential curve over the entire range if we are interested only in emission or in absorption shapes. Returning to Fig. 2, the absorption shape is determined primarily by the potential curves about points "a" and "b." Likewise the emission shape also is determined primarily from the curves about "c" and "d." Therefore, one would expect that the general features given in Sec. 9 will be observed. While the general features are given by the development, the calculations are not truly exact. The change in the frequency and higher-order terms must have a detectable influence. These corrections must somehow be used to explain the asymmetry at high temperatures. How to do this remains an unsolved problem.

In Secs. 10 and 18 several more general models are considered where various corrections are introduced in the expression for the moments. The effects on $\bar{\epsilon}$ and m^2 are, however, small and usually beyond the range of detection, except for the case where we have two radically different frequencies. We thus conclude that the basic equations have considerable validity. This field needs more reliable data and its complete analysis in terms of absorption and emission. In Sec. 11 some methods of analysis are suggested.

ACKNOWLEDGMENTS

Professor F. Seitz is thanked for suggesting that this review be written and for several helpful discussions. The assistance of Dr. A. Ellett in writing the manuscript is acknowledged. Further, the author is grateful to J. D. Konitzer for carrying out a detailed experimental investigation of the shape and thermal behavior of the F center in KCl. The author also thanks Mrs. B. Heckenbach for typing the manuscript.

Note added in proof.—Some interesting recent data have been reported by K. Maeda, J. Phys. Soc. Japan 14, 478 (1959).

13. APPENDIX I. LIST OF SYMBOLS

Although an attempt was made to associate a single physical concept with every symbol, this has not always been possible; thus, ϵ has been used for an eigenvalue as well as for a photon energy. Similarly, m has two distinct meanings. Usually subscripts have been employed to differentiate between physical concepts, or the symbols have been employed at distinct parts of the text. Gaussian units have been employed except for experimental values which are reported in electron volts. Only the important symbols are listed here.

a_s—Smakula's constant—ratio of the area under an absorption curve to the product of $\alpha_m H$; also $(G_m H)^{-1}$
b—local field correction (Sec. 8)
c—velocity of light
e—charge on the electron
f—oscillator strength (8.16)
g—shape factor associated with items (a) and (b) of Sec. 7, as well as the dispersion of the modes
$\hbar$—Planck constant divided by 2π
h_e—electronic Hamiltonian (3.3)
h_v—vibration Hamiltonian (3.4)
k—Boltzmann constant
(k)—the kth nuclei
m—See (7.3) except (8.16)
q_j—reduced coordinate of the jth normal mode (In Part II it is associated with the ground state.)
Δq_j—shift in the equilibrium position $q_j'-Q_j$ (3.24)
p—see (9.13)
p_t—see (9.22)
p_h—see (9.29)
p_j—momentum operator associated with q_j
$\mathbf{r}$—coordinate of all the electrons of interest to the problem
t—kinetic energy operator associated with the excess electron
v_j—vibration quantum number associated with the jth mode, v_j (ground), and v_j' (upper)
v_t—total vibrational quantum number, i.e., $\sum_j v_j$
$\bar{v}$—thermal average quantum number of single mode
B—reference (B2)
E—total energy in vibrational state v_t associated with the ground-electronic state
E'—total energy in vibrational state v_t' associated with the upper-electronic state

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

E^a—activation energy for absorption
E^e—activation energy for emission
E^t—thermal activation energy
G_A—see (9.19)
G_n—normalized shape factor associated with the absorption or emission of phonons
G_m—maximum value of G_n
H—width at half-height (half-width) (7.4)
H—total Hamiltonian [see (3.12), (3.13), (5.4), and (5.10) for various meanings]
K—see (10.1a)
LST—Lyddane-Sack-Teller relation (see B, p. 86)
M_n—nth moment of the absorption curve (7.1)
N—number of imperfections per unit volume
Q_j—reduced coordinate of the jth mode associated with the upper state
$P(v_t)$—see (8.14)
$\mathbf{R}$—coordinate of all the nuclei
$\mathbf{R}_L$—value of $\mathbf{R}$ which minimizes V_L
$\mathbf{R}_n$—see (3.4a)
S—Huang-Rhys factor $\frac{1}{2}\sum_j \epsilon_j^2/\hbar\omega_j^3$ (4.3a) and (10.23)
T—kinetic energy operator for all the nuclei
T'—T which only operates on the χ's
T_e—kinetic energy operator for all the electrons
V—total potential energy of the system
V_e—potential energy due to an excess electron
V_I—potential energy due to a point imperfection
V_L—potential energy of a perfect lattice
$X_\alpha(k)$—actual coordinate of the kth particle in the αth direction
α—absorption coefficient
α_m—absorption coefficient at ϵ_m
β_j—see (8.14b)
ϵ—stands for an eigenvalue
ϵ_g—ground state eigenvalue of h_e
ϵ_j—see (3.18a)
ϵ_{jk}—see (3.18b)
ϵ_n—nth eigenvalue of h_e
ϵ_u—upper state eigenvalue of h_e
ϵ_v—vibrational eigenvalue
$\Delta\epsilon(\mathbf{R})=\epsilon_u(\mathbf{R})-\epsilon_g(\mathbf{R})$—see after (5.3) for special meaning in Secs. 5(a) and 14
$\Delta\epsilon$ or $\Delta\epsilon(\theta)=\Delta\epsilon(\mathbf{R}_g)$ in Part II
ϵ—also stands for a photon energy
ϵ_m—photon energy for maximum absorption
$\bar{\epsilon}$—see (7.2)
θ—absolute temperature
ν—*angular* frequency of the absorbed light
ν_{ug}—see (9.3a)
σ—absorption cross section
φ—electronic eigenfunction [φ (ground) and φ' (upper) are also used]
χ_j and χ_j'—ground- and upper-state vibrational eigenfunctions associated with the jth mode
χ and χ'—total vibrational eigenfunctions of the ground and excited states
Ψ—total eigenfunction of the system
ω—the effective angular frequency of the interacting phonons
ω_j—angular frequency associated with the jth mode

14. APPENDIX II. MEYER-PEKAR METHOD

This appendix relates the method developed for the trapped electron (5a) to the techniques used previously. Meyer (M8) has made the most extensive calculation. We now split $H_0(A)$ of (5.2) as follows:

$$H_L=T'+V_L+V_I \tag{14.1}$$

$$H_e=t+V_e(\mathbf{R}_I) \tag{14.2}$$

and

$$H_{Le}=V_e(\mathbf{R})-V_e(\mathbf{R}_I). \tag{14.3}$$

The division of H_0 here is not identical to the one made previously. Equation (14.1) determines a set of normal modes, Λ_j, from an expansion about the point defined by

$$\left[\frac{\partial}{\partial X_\alpha(k)}(V_L+V_I)\right]_{\mathrm{RI}}=0. \tag{14.4}$$

There is a close similarity [see Eqs. (3.12a) and (3.12b)] between H_L+H_e and $H_0(S)$ and between H_{Le} and $H_1(S)$. The $H_0(S)$ and $H_1(S)$ here correspond to the impurity model of Sec. 5. The potential energy term in $h_v(\mathbf{R})$ has been modified slightly. Meyer refers to H_{Le} as the lattice-center interaction. This term is actually given by $H_1(A)$.

An appropriate way to find a solution is to expand H_{Le}, as follows,

$$H_{Le}=\sum_j a_j\Lambda_j+\tfrac{1}{2}\sum_{jk} b_{jk}\Lambda_j\Lambda_k. \tag{14.5}$$

We assume that the eigenfunctions φ_n and eigenvalues $\Delta\epsilon_n$ of H_e are known at $\mathbf{R}_I$. The perturbation method is now used to obtain eigenvalues and eigenfunctions of H_e+H_{Le} for any $\mathbf{R}$. Thus,

$$\varphi_n(\mathbf{R})=\varphi_n+\sum_j\sum_m{}'\frac{\langle\varphi_m|\,a_j\,|\varphi_n\rangle}{\Delta\epsilon_n-\Delta\epsilon_m}\varphi_m\Lambda_j \\ +\text{higher terms.} \tag{14.6}$$

In the last term $m\neq n$, and we assume that the eigenvalues associated with H_e are nondegenerate. The argument $\mathbf{R}_I$ of φ and $\Delta\epsilon$ on the right-hand side of (14.6) has been omitted. The higher-order terms which are proportional to $\Lambda_j\Lambda_k$ arise from the use of second-order perturbation theory as well as from the last term of (14.5).

The $\varphi_n(\mathbf{R})$'s must form an orthogonal set with respect to $\mathbf{r}$ for a fixed value of $\mathbf{R}$. This follows since the complete series (14.6) gives eigenfunctions of the Hamiltonian H_e+H_{Le}. This was not evident to Meyer.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

Associated with (14.6) is the eigenvalue

$$\Delta\epsilon_n(\mathbf{R})=\Delta\epsilon_n+\sum_j\langle\varphi_n|a_j|\varphi_n\rangle\Lambda_j$$
$$+\tfrac{1}{2}\sum_{jk}\langle\varphi_n|b_{jk}|\varphi_n\rangle\Lambda_j\Lambda_k$$
$$+\sum_m{}'\sum_{jk}\frac{1}{\Delta\epsilon_n-\Delta\epsilon_m}\langle\varphi_n|a_j|\varphi_m\rangle\langle\varphi_m|a_k|\varphi_n\rangle\Lambda_j\Lambda_k. \quad (14.7)$$

Terms of (14.7) may be identified with those which arise in an expansion about $\mathbf{R}_l$. We assume proper convergence. In such an expansion, terms similar to those in (3.18) appear. Thus,

$$\epsilon_j=\langle\varphi_n|a_j|\varphi_n\rangle \quad (14.8a)$$

and

$$\epsilon_{jk}=\langle\varphi_n|b_{jk}|\varphi_n\rangle$$
$$+\sum_m{}'\frac{1}{\Delta\epsilon_n-\Delta\epsilon_m}\langle\varphi_n|a_j|\varphi_m\rangle\langle\varphi_m|a_k|\varphi_n\rangle. \quad (14.8b)$$

Equation (14.8a) is a form of (3.16). The first term in (14.8b) was omitted by Meyer. One may use the ϵ_j's exactly as in Sec. 3 to determine the relation between the true q's and the Λ's. One should stress that H_{Le} does not arise from a true electron-phonon interaction but occurs only because a fictitious definition of the phonon field was used. On occasion the Λ's are assumed to have a very simple form, and the ϵ's obtained by these means correspond only very approximately with the real ones.

15. APPENDIX III. MEHLER'S FORMULA

Mehler's formula, Eq. (9.4), is "well known" in the theory of Hermite polynomials and has been used since it was discovered in 1866. Its proof is not available in standard texts. The steps are rather simple, providing one interchanges two integrals with an infinite sum. A proof due to Hardy is presented. It originally appears in a paper by Watson (W1). We shall not concern ourselves with the convergence of the series, so the proof as presented is not completely rigorous. A standard integral of interest is

$$\int_{-\infty}^{\infty}\exp[-a^2u^2+ibu]du=\frac{\sqrt{\pi}}{a}\exp[-b^2/4a^2]; \quad (15.1)$$

hence,

$$H_v=(-1)^v\exp[x^2]\frac{d^v}{dx^v}\exp[-x^2]$$
$$=\frac{(-2i)\exp[x^2]}{\sqrt{\pi}}\int_{-\infty}^{\infty}u^v\exp[-u^2+2ixu]du. \quad (15.2)$$

One would like to examine the sum

$$\sum_0^\infty t^v\chi_v(x)\chi_v(y)$$
$$=\sum_v\frac{1}{2^v v!\sqrt{\pi}}\exp\{-\tfrac{1}{2}(x^2+y^2)\}t^vH_v(x)H_v(y)$$
$$=\pi^{-\frac{1}{2}}\exp\{\tfrac{1}{2}(x^2+y^2)\}\sum_v\int\int\frac{(-2tuw)^v}{v!}$$
$$\times\exp\{-u^2-w^2+2ixu+2iyw\}dvdw. \quad (15.3)$$

For convenience, the usual $\omega/\hbar$'s which appear in the χ's have been omitted. v is the quantum state of χ_v. Hardy's device is to change the order of the sum and the integrals. Further, he replaces $\sum[(-2tuw)^v/v!]$ by $\exp\{-2tuw\}$. The convergence of (15.3) is such that this is permissible, provided $t^2<1$. Watson states this requirement but does not give details. One can see that (15.4) requires this relation. After this step, Eq. (15.1) is used twice, going from right to left. Thus,

$$\sum_v t^{v+\frac{1}{2}}\chi_v(x)\chi_v(y)$$
$$=t^{\frac{1}{2}}\pi^{-\frac{1}{2}}\exp\{\tfrac{1}{2}(x^2+y^2)\}$$
$$\times\int_{-\infty}^{\infty}\int\exp\{-u^2-2tuw-w^2+2ixu+2iyw\}dudw$$
$$=t^{\frac{1}{2}}\pi^{-1}\exp\{\tfrac{1}{2}(x^2-y^2)\}$$
$$\times\int\exp\{-(1-t^2)u^2+2i(x-yt)u\}du$$
$$=\pi^{-\frac{1}{2}}\left[\frac{t}{1-t^2}\right]^{\frac{1}{2}}\exp\left\{\frac{x^2-y^2}{2}-\frac{(x-yt)^2}{1-t^2}\right\}$$
$$=\pi^{-\frac{1}{2}}\left[\frac{t}{1-t^2}\right]^{\frac{1}{2}}\exp\left\{-\frac{(x^2+y^2)}{2}\left(\frac{1+t^2}{1-t^2}\right)+xy\frac{2t}{1-t^2}\right\}. \quad (15.4)$$

If we set $t=e^{\xi}(\xi>0)$, Eq. (15.4) gives (9.4) after some elementary transformations which use the properties of the hyperbolic functions.

16. APPENDIX IV. PEKARIAN CURVE

Properties of Pekarian curves, Eq. (9.21), are discussed here. This equation gives the relative probability for a transition from $v_t=0$ to any $v_t'=p$ (Eq. 9.15). Since the number of modes does not enter into the problem, we again use the CC model as an illustration. The Pekarian shape is not limited to this case but also applies to the Fröhlich and LM models. The only serious limitation is the use of Eq. (9.7).

Absorption shapes do not seem to be completely symmetric (K2a, p. 118), and one may inquire as to the symmetrical properties of the Pekarian curve. Consider first the case where $S=20$. The most probable transi-

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

TABLE VI. The symmetry properties of a Pekarian curve $S=20$.

$p=v_t'$	23	22	21	20	19	18	17	16
$\frac{S^p}{p!}\left(\frac{S^S}{S!}\right)^{-1}$	$\frac{(20)^3}{(21)(22)(23)}$	$\frac{(20)^2}{(21)(22)}$	$\frac{20}{21}$	1	1	$\frac{19}{20}$	$\frac{19(18)}{(20)^2}$	$\frac{(17)(18)(19)}{(20)^3}$
Transition probability	0.7529	0.8658	0.9524	1	1	0.9500	0.8550	0.7268

tions occur for $v_t'=S-1$ and S. If one assumes that the transition probability for these jumps is unity, then one may obtain the properties in Table VI. The transition probabilities do not lie symmetrically about the midpoint, which is very nearly at 19.5.

To stress the asymmetry of the problem, Fig. 9 is included. The left-hand side is similar to Fig. 6 except that the upper state has been shifted twice as much, and only the case where $v_t=0$ is considered. On the right-hand side $|\{\chi'|\chi\}|^2$ is plotted against v_t'. On the x axis v_t' increases in going from right to left, in agreement with experimental procedure. It is evident that $\epsilon_v-\epsilon_m\neq\epsilon_m-\epsilon_r$ so that the Pekarian curve is asymmetric.

In Fig. 10 three Pekarian curves for $S=3$, 10, and 20 are plotted. On the curves the ratio of $\epsilon_v-\epsilon_m$ to $\epsilon_m-\epsilon_r$ is noted. These values show that the curve is slightly lopsided for all values of S and resembles a double Gaussian. The violet side is indeed almost a perfect Gaussian.

To calculate various factors associated with a Pekarian curve, we use (9.17) and let $\theta=0$. The following formula is obtained:

$$|\{\chi_p'|\chi_0\}|^2=Ge^{-s}=(S^p/p!)e^{-s}. \tag{16.1}$$

0 and p are the total vibrational quantum numbers of the states χ and χ'. The term $1/\omega$ has not been included. It determines the width (or height) of the curve. Also, we set $\omega=\hbar\omega=1$. Using (10.9) and (10.24b) we note that,

$$\sum_{v_t'}|\{\chi'|\chi_0\}|^2=1 \tag{16.2}$$

and

$$m^2=S. \tag{16.3}$$

Since $M_0=1$ it is temperature independent [Eq. (10.9)]. Hence, we know that the relation

$$\sum_{v_t}P(v_t)\sum_{v_t'}|\{\chi'|\chi_0\}|^2=1 \tag{16.4}$$

holds for any θ. Now, since

$$\sum_{v_t}P(v_t)=1, \tag{16.5}$$

we conclude that

$$\sum_{v_t'}|\{\chi'_{v_t'}|\chi_{v_t}\}|^2=1 \tag{16.6}$$

for any v_t. This is a sum-rule which has one surprising feature—it does not involve the frequency. In atomic spectra, the sum-rules involve the transition frequency. This means that we may not accept the atomic sum-rules without further careful considerations.

By numerical calculation we obtain Table VII. Here

TABLE VII. Some properties of Pekarian curves.

S	p_l	$(e^{-s}G)_{\max}$	H	Area	$H^2/m^2=H^2/S$	$a_s(P)$
3	2.5	0.229	4.18	1	5.82	1.0
5	4.5	0.180	5.28	1.01	5.58	1.06
10	9.5	0.126	7.46	1.00	5.57	1.07
20	19.5	0.0892	10.56	1.01	5.58	1.07
30	29.5	0.0728	12.92	1.01	5.56	1.07

the lines of Fig. 9 have been replaced by the envelope curve. Column 3 is the value of the absorption at the peak (at p_l).†††† The fifth column is the area (planimeter) under the curves when the factor e^{-s} is included. This area might be slightly larger than unity because we are replacing a sum by a smooth curve. At $S-\frac{1}{2}$ the envelope is larger than the values at S or at $S-1$. To obtain the sixth column, use was made of (16.3), and the seventh is obtained from the definition of Smakula's constant found in Sec. 7, combined with (16.2). Columns 6 and 7 should be compared with the value found for the Gaussian shape, that is, 5.545 and 1.064. The Lorentzian value of a_s is 1.57. The deviation from the Gaussian values is very small. The variation for small values of S is probably real; however, we attach no importance to the slight deviation found in the last four rows. One may readily superimpose various Pekarian curves unless S is smaller than 10.

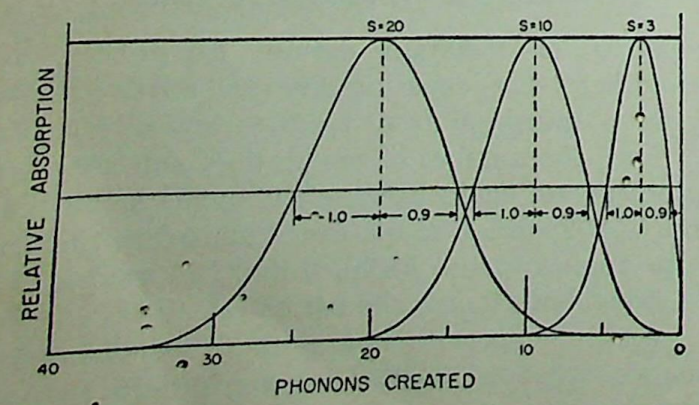


FIG. 10. Three Pekarian curves for various values of S. Note the lack of symmetry.

17. APPENDIX V. EFFECTS WHICH OCCUR WHEN f IS NOT A CONSTANT

We now consider what happens to the equations of Sec. 10 when the simplifying assumptions regarding f

†††† Since no fine structure is observed p will not be limited to integer values.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

are removed. One might use the Lax-Meyer method throughout, but the algebraic complexities become too involved. In Sec. 10 we replaced $\exp(iH_u t/\hbar)\times\exp(-iH_g t/\hbar)$ by $\exp[i(H_u-H_g)t/\hbar]$. It is correct only to terms in $(H_u-H_g)^2$. This arises because $\exp(iH_u t/\hbar)$ is only a shorthand notation for an infinite series and the terms in the two sums do not in general commute. Hence, this method is not used in the first section of this Appendix.

The two problems are: what happens to the expression for $\bar{\epsilon}$ and m^2 when f becomes a linear function of the frequency, or a linear function of the q's?

(a) Oscillator Strength Depends on the Frequency

First, we consider the effects of having the oscillator strength proportional to the frequency. For this purpose, the dipole matrix M is introduced.‡‡‡‡ It is assumed to be a constant. Further, we limit our considerations to the case where one may assume a single frequency and omit the ϵ_{jk}'s, i.e., the situation treated in Sec. 9.

Combining (8.16), (8.18), and (9.8) results in

$$\alpha(\nu)=bN\frac{4\pi^2}{3\hbar c}|\langle\varphi'|e\mathbf{r}|\varphi\rangle|^2\nu G(\nu)$$

$$=K'M^2\nu\frac{1}{2\pi}\int_{-\infty}^{\infty}e^{-i(\nu-\nu_{ug})t}$$

$$\times\exp\{-S[\coth\tfrac{1}{2}\beta-i\sin\omega t-\coth\tfrac{1}{2}\beta\cos\omega t]\}dt, \quad (17.1)$$

where

$$K'=4\pi^2bN/3\hbar c \quad (17.1a)$$

and

$$M=\langle\varphi'|e\mathbf{r}|\varphi\rangle. \quad (17.1b)$$

Since our interest is in the moments of the distribution, we assume that the absorption is a series of delta functions.

Taking the Fourier transformation of $\alpha(\nu)/\nu$, one arrives at

$$\int_{-\infty}^{\infty}[\alpha(\nu)/\nu]e^{i\nu t}d\nu=K'M^2e^{i\nu_{ug}t}$$

$$\times\exp\{-S[\coth\tfrac{1}{2}\beta-i\sin\omega t-\coth\tfrac{1}{2}\beta\cos\omega t]\}. \quad (17.2)$$

From this it follows that

$$M_0=(\hbar/i)\left\{(\partial/\partial t)\int_{\infty}^{\infty}[\alpha(\nu)/\nu]e^{i\nu t}d\nu\right\}_{t=0}$$

$$=\hbar K'M^2\{\nu_{ug}+S\omega\}=K'M^2\Delta\epsilon=K\hbar\bar{f}, \quad (17.3)$$

where

$$\bar{f}=2m\bar{\nu}M^2/3e^2\hbar \quad (17.3a)$$

and

$$\bar{\nu}=\Delta\epsilon/\hbar. \quad (17.3b)$$

Use has been made of (10.1a) and (10.14). (17.3) should be compared to (10.9). Further,

$$M_1=-\hbar^2\{(\partial^2/\partial t^2)\textstyle\int_{-\infty}^{\infty}[\alpha(\nu)/\nu]e^{i\nu t}d\nu\}_{t=0}$$
$$=K'M^2[(\Delta\epsilon)^2+\hbar^2\omega^2S\coth\tfrac{1}{2}\beta]$$
$$=K'M^2\Delta\epsilon[\Delta\epsilon+(\hbar^2\omega^2/\Delta\epsilon)S\coth\tfrac{1}{2}\beta] \quad (17.4)$$

and

$$M_2=-(\hbar^3/i)\{\partial^3/\partial t^3)\textstyle\int_{-\infty}^{\infty}[\alpha(\nu)/\nu]e^{i\nu t}d\nu\}_{t=0}$$
$$=K'M^2\Delta\epsilon[(\Delta\epsilon)^2+3\omega^2\hbar^2S\coth\tfrac{1}{2}\beta$$
$$+(\omega^3\hbar^3/\Delta\epsilon)S]. \quad (17.5)$$

Further,

$$\bar{\epsilon}=\Delta\epsilon+(\hbar^2\omega^2/\Delta\epsilon)S\coth\tfrac{1}{2}\beta \quad (17.6)$$

and

$$m^2=\omega^2\hbar^2S\{\coth\tfrac{1}{2}\beta+(\omega\hbar/\Delta\epsilon)$$
$$-[\hbar^2\omega^2/(\Delta\epsilon)^2]S\coth^2\tfrac{1}{2}\beta\}. \quad (17.7)$$

Equation (17.6) shows that $\bar{\epsilon}$ shifts to the *violet* with rising temperatures. The shift is very small, since $\hbar^2\omega^2S/\Delta\epsilon$ is of the order of 0.0015 ev. If $\hbar\omega$ were 0.08 ev and $S\approx30$ this term would be observable. This might occur in LiF if the transverse optical mode were important for the interaction. m^2 now has a constant term and a term proportional to $\coth^2\frac{1}{2}\beta$. The negative sign on this term means that m^2 can be negative at *very* high temperatures. This anomalous situation arises from the mathematical formulation and has nothing to do with actual emission. By returning to (17.1) we see that $\alpha(\nu)<0$ if $\nu<0$. Since the mathematical techniques require integration from $-\infty$ to $+\infty$, the integral $\int_{-\infty}^{\infty}(\nu-\bar{\nu})^2\nu G(\nu)d\nu$ can be negative in spite of the fact that $(\nu-\bar{\nu})^2>0$. The author expects that both additional terms in (17.7) are very small and will not be detected experimentally.

The form (17.7) is not valid at extremely high temperatures, say 5000°K, if $(\hbar\omega)^2S/(\Delta\epsilon)^2$ is of the order of 10^{-3}. For the F center in KCl this term is of the order of 8×10^{-4}, and our interests are limited to temperatures below 600°K. The above calculations are most probably valid under these conditions. They indicate that the assumption that f is independent of ν does not have an important influence on the moments. In the language of Dexter (D1, p. 392), we have shown when the "narrow-band approximation" is valid.

(b) Oscillator Strength Depends on the q's

In the case where $f^{\frac{1}{2}}$ depends on the normal modes, the solution is straightforward but extremely laborious. To simplify the problem somewhat the CC model will be used, thus $f^{\frac{1}{2}}$ is taken in the form

$$f^{\frac{1}{2}}=f_0^{\frac{1}{2}}+f_1^{\frac{1}{2}}q. \quad (17.8)$$

Equation (17.8) is a possible assumption, and it is

‡‡‡‡ For simplicity of notation M is considered to be a real scalar.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

made without any profound theoretical justification? In (17.8) q is the coordinate associated with the ground state. The normal mode for the excited state is taken in the form $q-\Delta q=Q$. Further, we again assume that the ϵ_{jk}'s are very small so that the frequencies in both states are equal. As before, $f^{\frac{1}{2}}$ is taken as real; further, our considerations will be limited to terms in $f_1^{\frac{1}{2}}$. Equation (10.6) cannot be used, and we require some more complex equations given by Lax (L1), namely,

$$M_0=\hbar K \operatorname{Av}\{\chi|f|\chi\} \tag{17.9}$$

$$M_1=\hbar K \operatorname{Av}\{\chi|f^{\frac{1}{2}}H_u f^{\frac{1}{2}}-fH_g|\chi\} \tag{17.10}$$

and

$$M_2=\hbar K \operatorname{Av}\{\chi|f^{\frac{1}{2}}H_u^2 f^{\frac{1}{2}} -2f^{\frac{1}{2}}H_u f^{\frac{1}{2}}H_g+fH_g^2|\chi\}. \tag{17.11}$$

Equation (17.10) is equivalent to (10.12) provided

$$\{\chi|H_uH_g|\chi\}=\{\chi|H_gH_u|\chi\} \tag{17.12}$$

when f is independent of q.

To establish (17.12) we require some of the properties of the χ's. One may show (M7, p. 121 and 358)§§§§

$$(\partial/\partial q)\chi_v=(\omega/\hbar)^{\frac{1}{2}}\{(\tfrac{1}{2}v)^{\frac{1}{2}}\chi_{v-1}-[\tfrac{1}{2}(v+1)]^{\frac{1}{2}}\chi_{v+1}\} \tag{17.13}$$

$$q\chi_v=(\hbar/\omega)^{\frac{1}{2}}\{[\tfrac{1}{2}(v+1)]^{\frac{1}{2}}\chi_{v+1}+(v/2)^{\frac{1}{2}}\chi_{v-1}\}. \tag{17.14}$$

Hence

$$\{\chi|q(\partial/\partial q)|\chi\}=-\tfrac{1}{2} \tag{17.15}$$

and

$$\{\chi|TV_{u,g}|\chi\}=\{\chi|V_{u,g}T|\chi\}. \tag{17.16}$$

To establish (17.16) we have used (3.26) but placed no restriction on the v's; hence, (17.16) applies to the case considered in Sec. 10 and is the justification for Eq. (10.12). A similar argument has not been found for terms in $(H_u-H_g)^3$, and the technique of Sec. 10 requires lengthy and complex algebra.

Returning to (17.8), we note from (17.13) and (17.14) that our interest is limited to even terms of the q's, i.e., such as q^2 or $q(\partial/\partial q)$, not q^3 or $\partial/\partial q$. Hence, substituting (17.8) in (17.9), one obtains

$$M_0=\hbar K f_0. \tag{17.17}$$

Had terms in f_1 been included, we would have found that M_0 is temperature *dependent*. Now, substituting into (17.10), we obtain

$$M_1=\hbar K \operatorname{Av}\{\chi|f_0(V_u-V_g) -2f_0^{\frac{1}{2}}f_1^{\frac{1}{2}}\omega^2(\Delta q)q^2|\chi\}, \tag{17.18}$$

$$=\hbar K(f_0\Delta\epsilon-\omega\hbar\Delta q f_0^{\frac{1}{2}}f_1^{\frac{1}{2}}\coth\tfrac{1}{2}\beta) \tag{17.19}$$

$$=\hbar K f_0\left\{\Delta\epsilon-\omega\hbar\left[\frac{M(d)-M(a)}{M(a)}\right]\coth\tfrac{1}{2}\beta\right\}. \tag{17.19a}$$

$M(a)$ and $M(d)$ are dipole moments at positions (a) and (d) of Fig. 2. The ratio $[M(d)-M(a)]/M(a)$ must be of the order of 0.1 or much less, probably much less. The second term of (17.19a) cannot possibly be larger than 0.001, and we would not expect to detect it experimentally. Further,

$$M_2=\hbar K \operatorname{Av}\{\chi|f_0(V_u-V_g)^2+f_0^{\frac{1}{2}}f_1^{\frac{1}{2}}[TV_uq+V_uTq +qTV_u-3qV_uT+2q(V_u-V_g)^2]|\chi\}. \tag{17.20}$$

Since

$$\{\chi|TV_uq|\chi\}=\{\chi|qV_uT|\chi\} \tag{17.21a}$$

$$\{\chi|V_uTq|\chi\}=\{\chi|qTV_u|\chi\} =\{\chi|qV_uT|\chi\}-(\hbar^2/2)\omega^2\Delta q, \tag{17.21b}$$

we have

$$M_2=\hbar K \operatorname{Av}\{\chi|(V_u-V_g)^2f_0+f_0^{\frac{1}{2}}f_1^{\frac{1}{2}}[2q(V_u-V_g)^2 -\hbar^2\omega^2\Delta q]|\chi\}. \tag{17.22}$$

The third term gives rise to a temperature independent factor in m^2. Taking the indicated average in (17.22) results in

$$M_2=\hbar K\{f_0[(\Delta\epsilon)^2+\hbar^2\omega^2S\coth\tfrac{1}{2}\beta] -\Delta q f_0^{\frac{1}{2}}f_1^{\frac{1}{2}}[2\Delta\epsilon\hbar\omega\coth\tfrac{1}{2}\beta+\hbar^2\omega^2]\}. \tag{17.23}$$

Returning to (7.3) we obtain

$$m^2=\hbar^2\omega^2[S\coth\tfrac{1}{2}\beta-\Delta q(f_1^{\frac{1}{2}}/f_0^{\frac{1}{2}})]. \tag{17.24}$$

To estimate the magnitude of the second term we again use Fig. 2. From the definitions of the f's and Δq one obtains

$$\Delta q f_1^{\frac{1}{2}}/f_0^{\frac{1}{2}}=[M(d)-M(a)]/M(a). \tag{17.25}$$

The maximum correction term that one may expect in (17.24) is 0.1. Such small values could barely be detected experimentally (Sec. 10). Meyer (M8) suggested that this term equals -10. Such large values cannot be explained readily by the theory.

The conclusion is rather straightforward. The assumptions used in arriving at (8.18) are most probably oversimplified. However, they do not produce major changes in the moments of the distribution or in the shapes. Some corrections to the moments have been worked out here. They are certainly too small to be measured, and the author cannot envisage any theoretical calculation which will be sufficiently accurate to determine them. One may show experimentally, however, that M_3 has a complex temperature dependence (K4). The theoretical reason for this is at present unknown. The author believes this will come out of a more realistic model.

§§§§ The subscript is the vibrational quantum number.

BIBLIOGRAPHY

A1. M. Abraham and R. Becker, *The Classical Theory of Electricity and Magnetism* (Blackie and Son, Ltd., London, 1950).
B1. R. L. Bjork, Phys. Rev. **105**, 456 (1957).
B2. M. Born and K. Huang, *Dynamical Theory of Crystal Lattices* (Clarendon Press, Oxford, 1954).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

B3. M. Born and J. Oppenheimer, Ann. phys. **84**, 457 (1927).
B4. M. A. Bredig, Z. Physik **46**, 73 (1927).
B5. E. Burstein and J. J. Oberly, Natl. Bur. Standards (U.S.) Circ. No. 519, 285 (1952).
C1. T. G. Castner and W. Kanzig, J. Phys. Chem. Solids **3**, 178 (1957).
C2. D. W. Compton (private communication).
C3. D. W. Compton and C. C. Klick, Phys. Rev. **110**, 349 (1958).
D1. D. L. Dexter in *Solid State Physics*, edited by F. Seitz and D. Turnbull (Academic Press, Inc., New York, 1958), Vol. 6.
D2. W. H. Duerig and J. J. Markham, Phys. Rev. **88**, 1043 (1952).
D3. C. Z. van Doorn and Y. Haven, Philips Research Repts. **11**, 479 (1956).
D4. D. Dutton and R. Maurer, Phys. Rev. **90**, 126 (1953).
F1. H. Frölich, Proc. Roy. Soc. (London) **A160**, 230 (1937).
F2. H. Frölich, Advances in Phys. **3**, 325 (1954).
F3. Fröhlich, Pelzer, and Zienau, Phil. Mag. **41**, 221 (1950).
H1. H. Haken, Z. Physik, **135**, 408 (1953).
H2. A. Haug, Z. Physik **146**, 75 (1956).
H3. W. Heitler, *The Quantum Theory of Radiation* (Clarendon Press, Oxford, 1954), third edition.
H3a. F. A. Henglein, Z. physik Chem. **115**, 91 (1925).
H4. K. Huang, Proc. Roy. Soc. (London) **A208**, 352 (1951).
H5. K. Huang and A. Rhys, Proc. Roy. Soc. (London) **A204**, 406 (1950).
I1. H. F. Ivey, Phys. Rev. **72**, 341 (1947).
J1. H. Jeffreys and B. S. Jeffreys, *Methods of Mathematical Physics* (University Press, Cambridge, England, 1946).
K1. H. Kanzaki, Phys. Rev. **110**, 1063 (1958).
K2. C. C. Klick, Phys. Rev. **85**, 154 (1952).
K2a. C. C. Klick and J. H. Schulman, in *Solid State Physics*, edited by F. Seitz and D. Turnbull (Academic Press, Inc., New York, 1957), Vol. 5.
K3. J. D. Konitzer and J. J. Markham, Phys. Rev. **107**, 685 (1957).
K4. J. D. Konitzer and J. J. Markham (to be submitted).
K5. M. A. Krivoglaz and S. I. Pekar, Trudy Inst. Fiz., Akad. Nauk Ukr. S. S. R. **4**, 37 (1953) (in Russian).
K6. R. Kubo and Y. Toyozawa, Progr. Theoret. Phys. (Kyoto) **13**, 160 (1955).
L1. M. Lax, J. Chem. Phys. **20**, 1752 (1952).
L2. M. Lax, *Photoconductivity Conference*, edited by R. G. Breckenridge (John Wiley and Sons, Inc., New York, 1956), p. 111.
L2a. M. Lax and E. Burstein, Phys. Rev. **100**, 592 (1955).
L3. K. Lonsdale, *Crystals and X-rays* (G. Bell and Sons, London, 1948).
M1. Mador, Markham, and Platt, Phys. Rev. **91**, 1277 (1953).
M2. Maradudin, Mazur, Montroll, and Weiss, Revs. Modern Phys. **30**, 175 (1958).
M3. J. J. Markham, J. Chem. Phys. **16**, 580 (1948).
M4. J. J. Markham, Phys. Rev. **103**, 588 (1956).
M5. J. J. Markham and J. D. Konitzer, J. Chem. Phys. **29**, 873 (1958); **30**, 328 (1959).
M6. J. J. Markham and F. Seitz, Phys. Rev. **74**, 1014 (1948).
M7. H. Margenau and G. M. Murphy, *The Mathematics of Physics and Chemistry* (D. Van Nostrand and Company, Inc., Princeton, New Jersey, 1956), second edition.
M8. H. J. G. Meyer, Physica **21**, 253 (1955).
M9. H. J. G. Meyer, Halbleitprobleme **3**, 230 (1956).
M10. E. Mollwo, Z. Physik **85**, 56 (1933).
M11. J. P. Molnar, "The absorption spectra of trapped electrons in alkali halide crystals," thesis, Massachusetts Institute of Technology, 1940 (unpublished).
M12. N. F. Mott and R. W. Gurney, *Electronic Processes in Ionic Crystals* (Clarendon Press, Oxford, 1948).
M13. T. Muto, Progr. Theoret. Phys. (Kyoto) **4**, 181 (1949).
O1. R. C. O'Rourke, Phys. Rev. **91**, 265 (1953).
P1. S. I. Pekar, J. Exptl. Theoret. Phys. (U.S.S.R.) **20**, 510 (1950), translated by Professor B. Goodman.
P2. S. I. Pekar, J. Exptl. and Theoret. Phys. (U.S.S.R.) **22**, 641 (1952), translated by A. Pingell, Naval Research Laboratory.
P3. S. I. Pekar, Uspekhi Fiz. Nauk **50**, 197 (1953), translated by M. F. Friedman of Newtonville, Massachusetts.
P4. S. I. Pekar, *Untersuchungen über die Elektronentheorie der Kristalle* (Akademische-Verlagsgesellschaft, Berlin, 1954).
P5. P. Pringsheim, Z. Physik **136**, 573 (1954).
R1. G. A. Russell and C. C. Klick, Phys. Rev. **101**, 1473 (1956).
S1. F. Seitz, Revs. Modern Phys. **26**, 7 (1954).
S2. F. Seitz, *The Modern Theory of Solids* (McGraw-Hill Book Company, Inc., New York, 1940).
S3. J. H. Simpson, Proc. Roy. Soc. (London) **A197**, 266 (1949).
S4. J. C. Slater, *Quantum Theory of Matter* (McGraw-Hill Book Company, Inc., New York, 1951).
S5. J. C. Slater, *Introduction to Chemical Physics* (McGraw-Hill Book Company, Inc., New York, 1939).
S6. W. A. Smith, "Energy level calculations for *F* centers and positive ions in alkali halides," Knolls Atomic Power Laboratory (Rept. KAPL-1720, April 1, 1957).
V1. J. H. Van Vleck, *The Theory of Electric and Magnetic Susceptibilities* (Clarendon Press, Oxford, 1932).
V2. H. D. Vasileff, Phys. Rev. **96**, 603 (1954).
W1. G. N. Watson, J. London Math. Soc. **8**, 194 (1933).
W2. F. E. Williams, J. Chem. Phys. **19**, 457 (1951).
W3. F. E. Williams and M. H. Hebb, Phys. Rev. **84**, 1181 (1951).
W4. A. H. Wilson, *The Theory of Metals* (University Press, Cambridge, England, 1953).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

Energy Transport in Organic Phosphors

S. C. GANGULY AND N. K. CHAUDHURY

Solid Physics Laboratory, Jadavpur University, Calcutta, India

CONTENTS

1. Introduction 990
(a) Photon Absorption and Reabsorption 990
(b) Collision Process 992
Electric Dipole-Dipole Interaction 993
Electric Dipole-Quadrupole Interaction . . . 994
Exchange Effects and Other Forbidden Transitions 995
Electric Dipole-Magnetic Dipole Interaction 996
Host Sensitization 996
(c) Exciton Migration 996
(d) Electron and Hole Migration 999
2. Fluorescence Transfer in Organic Mixed Crystals 999
(a) Exciton and Absorption 999
3. Electronic States of Aromatic Crystals 1002
(a) Benzene Crystal 1002
(b) Naphthalene Crystal 1003
(c) Anthracene Crystal 1004
(d) Naphthacene Crystal 1005
(e) Phenanthrene Crystal 1005
(f) Hexamethyl Benzene Crystal 1005
4. Influence of Crystal Size on the Spectral Nature and Decay Time of Fluorescence 1006
5. Scintillation from Organic Crystals 1008
6. Photoconductivity of Organic Crystals 1010
7. Some New Results 1015
8. Summary 1016
9. Acknowledgments 1016

1. INTRODUCTION

THE mechanism of energy transfer from one constituent to the other in solids, solutions, and gaseous mixtures has created considerable interest during recent years. Until very recently these discussions[1,2] were confined mostly to the inorganic phosphors in general, and alkali halides in particular. Along with the inorganic atoms and molecules, more than a few experimental results on benzene and benzene derivatives are also reported in the literature.[3–6] For instance, the fluorescence of indigo vapor is not excited by light of wavelength between 2700 to 2400 A, but in the presence of aniline vapor it gives violet indigo fluorescence. The incident light is first absorbed by the aniline vapor and then the absorbed energy is transferred to the indigo vapor. Similarly, aniline in presence of benzene vapor gives strong fluorescence under weak excitation. Zinc radiation at 2500 A produces only a very weak fluorescence in aniline vapor extending from 2850 to 2500 A. In a mixture of aniline and benzene vapor the benzene fluorescence is weak though it is easily excited by primary light of wavelength 2500 A, while the aniline fluoresces strongly. Apparently, the transfer of excitation energy from the benzene to the aniline molecules has high efficiency. These are instances of sensitized fluorescence which are commonly described as the process whereby an impurity atom or molecule (activator) devoid of any appreciable absorption band in a given spectral region is made to radiate upon absorption of excitation energy in this region by the host lattice or an impurity atom or molecule (sensitizer) embedded in it. In all cases of sensitization, the problem which is of the most fundamental significance is as follows: What is the mechanism of propagation of the excitation energy from the sensitizer to the activator and its ultimate transfer? Where the sensitizer and the activator are close to each other, the excitation energy does not have to move far and the processes of propagation and transfer both take place at nearly the same position in the lattice. There are also cases of sensitized luminescence where absorption of the exciting energy takes place all over the matrix lattice while the luminescence transitions occur in or near the few activators. These cases are characterized by propagation of energy through the lattice over a considerable distance before transfer of the excitation energy can occur. There are four distinct processes by which transfer of the excitation energy can take place. They are (a) photon absorption and reabsorption, known as photon cascade; (b) collision; (c) exciton migration; and (d) electron and hole migration. We discuss them briefly, bringing out the conditions to which each is subject. Particular emphasis is given to the discussion of anthracene-naphthacene and anthracene-naphthalene systems in crystals.

(a) Photon Absorption and Reabsorption

This mechanism of energy transfer, also called the photon cascade process, becomes of quantitative importance in a system if there exists strong overlap between the emission and the absorption spectra and

[1] F. A. Kroger, *Some Aspects of the Luminescence of Solids* (Elsevier Publishing Company, Inc., Houston, Texas, 1948).
[2] H. W. Leverentz, *Introduction to Luminescence of Solids* (John Wiley & Sons, Inc., New York, 1950).
[3] P. Pringsheim, *Fluorescence and Phosphorescence* (Interscience Publishers, Inc., New York, 1950).
[4] M. Furst and H. Kallmann, Phys. Rev. 94, 503 (1954).
[5] M. Furst and H. Kallmann, Phys. Rev. 97, 583 (1955).
[6] M. Furst and H. Kallmann, J. Chem. Phys. 23, 607 (1955).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

if the system has a high absorption coefficient in the area of overlap, comparable with that at the wavelength of excitation. The system may be a pure one containing no foreign constituents, the region of overlap being between the emission and the absorption spectra of the constituents of the pure system. If the system is an impure one, containing impurity atoms or molecules embedded in it, the region of overlap may be provided by the emission spectrum of the pure system and the absorption spectrum of impurity constituents. If we consider such a crystal system, the part of the primary molecular fluorescence that is emitted in the region of overlap, is practically reabsorbed by the crystal within a small fraction of its thickness. This reabsorbed radiation is re-emitted as fluorescence by the molecules which absorbed it. A fraction of it suffers reabsorption by succeeding layers of molecules, these processes of emission and reabsorption follow each other until all the excitation energy is either transformed into fluorescence outside the region of overlap so as to escape from the crystal as technical fluorescence or is thermally dissipated by internal conversion.

A quantitative formulation of the mechanism was developed by Birks[7] to explain the photofluorescent and scintillation properties of organic materials. Let us consider a mixed crystal containing solute molecules A immersed in solvent molecules S. On the photon cascade theory, the processes shown in Table I occur in the crystal, assuming that the exciting radiation is absorbed totally by S and that S does not absorb the fluorescence radiation of A.

When only processes 2 and 3 (see Table I) compete, there results the molecular fluorescence of S. Its quantum efficiency η_{0s} is given by

$$\eta_{0s}=\frac{1}{1+(K_{is}/K_{fs})}. \tag{1}$$

When the competing processes are described by 2, 3, 4, 5, and 9, the primary fluorescence of S is reabsorbed and re-emitted and only a part of the primary fluorescence S escapes from the crystal as technical fluorescence. The quantum efficiency η_{es} of this is given by

$$\eta_{es}=\frac{\eta_{0s}}{1+(K_S/K_{es})+(K_A/K_{es})[A]}+\frac{K_S}{K_{es}}\left\{\frac{\eta_{0s}}{1+(K_S/K_{es})+(K_A/K_{es})[A]}\right\}^2+\left(\frac{K_S}{K_{es}}\right)^2\left\{\frac{\eta_{0s}}{1+(K_S/K_{es})+(K_A/K_{es})[A]}\right\}^3=\frac{\eta_{0s}}{1+(1-\eta_{0s})(K_S/K_{es})+(K_A/K_{es})[A]}. \tag{2}$$

[7] J. B. Birks, Phys. Rev. 94, 1957 (1954).

TABLE I. Photon transfer process.

Process	Relative ratio[a]	Description of process
1. $S+h\nu \rightarrow S^*$	1	Excitation by light
2. $S^* \rightarrow S+h\nu_s$	$K_{fs}[S^*]$	Fluorescence of S
3. $S^* \rightarrow S$	$K_{is}[S^*]$	Internal quenching of S
4. $S+h\nu_s \rightarrow S^*$	$K_sK_{fs}[S^*]$	Self-absorption of S
5. $A+h\nu_s \rightarrow A^*$	$K_AK_{fs}[S^*][A]$	Absorption of A
6. $A^* \rightarrow A+h\nu_A$	$K_{fA}[A^*]$	Fluorescence of A
7. $A^* \rightarrow A$	$K_{iA}[A^*]$	Internal quenching of A
8. $A+h\nu_A \rightarrow A^*$	$K_AK_{fA}[A^*][A]$	Self-absorption of A
9. $h\nu_s \rightarrow$	$K_{es}K_{fs}[S^*]$	Escape of $h\nu_s$
10. $h\nu_A \rightarrow$	$K_{eA}K_{fA}[A^*]$	Escape of $h\nu_a$

[a] Here Brackets stand for concentrations in molecules per solvent molecule, S^* and A^* for excited molecules of S and A, and $K=(1/t)$ for the probability of a process of decay time t, i.e., number of events describing the process, that happen per unit time.

In the absence of internal quenching $K_{is}=0$ and $\eta_{0s}=\eta_{0A}=1$, so that (2) is reduced

$$\eta_{es}=\frac{1}{1+(K_A[A]/K_{es})}. \tag{3}$$

The quantum yield for transfer, i.e., the ratio of the number of photons transferred per second from S to A to the number of photon absorbed by S per second, neglecting nonradiative transactions in S becomes, in the case of molecular fluorescence,

$$\eta_{0T}=K_A[A]/K_S. \tag{4}$$

The molecular fluorescence quantum efficiency η_{0A} of A is given by

$$\eta_{0A}=\frac{1}{1+(K_{iA}/K_{fA})}. \tag{5}$$

The quantum efficiency (η_{eA}) of escape of (A) technical fluorescence is given by

$$\eta_{eA}=\frac{\eta_{es}\eta_{0A}(K_A/K_{es})[A]}{1+(1-\eta_{0A})(K_A'/K_{eA})[A]}. \tag{6}$$

In the absence of internal quenching, $K_{iA}=0$ and $\eta_{0s}=\eta_{0A}=1$ so that we have for (6)

$$\eta_{eA}=\eta_{es}(K_A/K_{es})[A]=1-\eta_{es}=\frac{(K_A/K_{es})[A]}{1+(K_A/K_{es})[A]}. \tag{7}$$

Here the quantities K_{es}, K_S, and $K_A(A)$ are interrelated by

$$K_{es}+K_S+K_A(A)=1.$$

They are dependent upon the relative absorption coefficients of S and A for the fluorescence of S and on the crystal size.

In pure crystal phosphors, processes 1, 2, 3, 4, and 9 are in operation. The molecular quantum efficiency (suffix S is now dropped) remains unchanged while the

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

technical quantum efficiency η_e takes the form

$$\eta_e = \frac{\eta_0}{1+(1-\eta_0)(K/K_e)} = \frac{1}{1+(K_i/K_eK_f)}; \quad [\because \; K+K_e=1]. \tag{8}$$

If internal quenching is not present, the molecular photofluorescence decay time $(t_f)_{00}=1/K_f$; internal quenching, if present, reduces it to $(t_f)_0$ given by

$$(t_f)=\frac{(t_f)_{00}}{1+[(t_f)_{00}/t_i]}=\eta_0(t_f)_{00}. \tag{9}$$

In case of a large thick crystal, stepwise self-absorption lengthens the duration of the interval between the moment of primary excitation and the moment of escape of fluorescence so that the technical photofluorescence decay time t_f is increased to

$$t_f=\frac{(t_f)_{00}}{K_e+[(t_f)_{00}/t_i]}. \tag{10}$$

When internal quenching is small enough to make $(t_f)_{00} \ll t$; (8) and (10) may be expressed approximately as follows:

$$t_f=\eta(t_f)_0 \tag{11}$$

and

$$\eta_e=(\eta_0)^N. \tag{12}$$

Here $N=1/K_e$ represents the average number of steps of molecular emissions passed before escape of technical fluorescence. The molecular and technical photofluorescence quantities $(t_f)_0$ and (t_f) can be obtained from measurements on microcrystalline and thick crystal specimens and thus N can be determined. Comparison of the molecular and the technical photofluorescence spectra leads to a determination of N.

(b) Collision Process

This mechanism was proposed by Cario and Franck[8] and to account for the disappearance of fluorescence of a gas due to the presence of foreign gas molecules. In this process when an excited atom or molecule collides during the lifetime of its excited state with a slow electron, it may give away its excitational energy to the electron, without any radiation, the products of the collision being an electron with enhanced kinetic energy and an unexcited atom or molecule. When the collision partner is a normal atom or molecule, the excited atom or molecule may give away a quantum of energy (E_2-E_1) to the unexcited one and the latter may take up the energy either as energies of translation, excitation or both; there being no loss of energy during the process as radiation. Such radiationless transfer of excitation energy is called a collision of the second kind, as distinguished from a collision of the first kind in which excitation is produced. The mechanism, demonstrated in gases by Franck, Cario, and many others, has been shown to take place also in some organic solutions.[3,9]

The quantum-mechanical theories of this transfer process as formulated by Perrin[10] and Forster[11], are based on resonance between allowed electric dipole transitions in the sensitizer and the activator. The resonance theory is extended by Dexter[12] to include cases of energy transfer in which an activator having a suitable emission spectrum can be made to fluoresce with energy transferred from an absorbing sensitizer despite the fact that the activator is forbidden direct absorption of energy. The transfer mechanisms of importance are by overlap of the electric dipole fields of the sensitizer and the activator, overlap of the dipole field of the former and the quadrupole field of the latter, and exchange effects, producing, respectively, sensitization of about 10^3-10^4, about 10^2, and about 30 lattice sites in a crystal surrounding each sensitizer. Dexter's theory is applicable to impurity sensitization both in solid and liquid systems. He adopts the physical model of a crystal in which the sensitizing impurity atoms or ions, of atomic concentration x_s, and the activators of atomic concentration x_a, are randomly arranged at suitable lattice sites, without mutual interaction. The time-dependent perturbation method of quantum mechanics gives the probability of transition of a system from its initial to final state as

$$(2\pi/h)|\langle H_1\rangle|^2\rho_E, \tag{13}$$

where $\hbar$ is Planck's constant divided by 2π, H_1, the matrix element of the perturbation to the Hamiltonian and ρ_E the density of final states. In this case the initial state ψ_1 is identified with the configuration of the system in which the sensitizer S is in the excited state ψ_s' while the activator A is in its ground state ψ_a and the final state ψ_2 is the configuration in which S is in its ground while A is excited to the state ψ_a'. Thus the probability that the energy of excitation is transferred from a particular sensitizer S to a particular activator A has the form,

$$P_{sa}=(2\pi/h)\rho_E\left|\int\psi_1^*H_1\psi_2 d\tau\right|^2, \tag{14}$$

as long as the states ψ_1 and ψ_2 are of the same energy.

To take account of the lack of definition of the initial and final levels of S or A caused by lattice vibrations, the wave functions are normalized on an energy scale and the density of states factor ρ_E is included within the

[8] G. Cario and J. Franck, Z. Physik 17, 202 (1928).

[9] J. Franck and R. Livingston, Revs. Modern Phys. 21, 505 (1949).

[10] F. Perrin, J. phys. radium 7, 1 (1936).

[11] Th. Forster, Ann. phys. 2, 55 (1948); Z. Electrochem. 53, 93 (1949).

[12] D. L. Dexter, J. Chem. Phys. 21, 836 (1953).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

normalization parameters. It is convenient to normalize the initial states of S and A (ψ_s' and ψ_a) and to express by the functions $p_s'(w_s')$ and $p_a(w_a)$ the probabilities that S is in the particular energy state denoted by w_s' and that A is in the state w_a so that

$$\int |\psi_s'(w_s')|^2 d\tau = \int |\psi_a(w_a)|^2 d\tau,$$
$$\int_0^{\alpha} p_s'(w_s')dw_s' = \int p_a(w_a)dw_a = 1. \tag{15}$$

The final states of S and A (ψ_s and ψ_a') are normalized on the energy scale by the following:

$$\frac{1}{\Delta w}\int_w^{w+dw} dw_s \int |\psi_s(w_s)|^2 d\tau = 1$$
$$\frac{1}{\Delta w}\int_w^{w+\Delta w} dw_a' \int |\psi_a'(w_a')|^2 d\tau = 1. \tag{16}$$

The properly antisymmetrized initial and final states wave functions describing S and A can now be written as

$$\psi_1(w_s', w_a) = \frac{1}{\sqrt{2}}\begin{vmatrix} \psi_s(\bar{r}_1, w_s') & \psi_s'(\bar{r}_2, w_s') \\ \psi_a(\bar{r}_1, w_a) & \psi_a(\bar{r}_2, w_a) \end{vmatrix}$$
$$\psi_2(w_s, w_a') = \frac{1}{\sqrt{2}}\begin{vmatrix} \psi_s(\bar{r}_1, w_s) & \psi_s(\bar{r}_2, w_s) \\ \psi_a'(\bar{r}_1, w_a') & \psi_a'(\bar{r}_2, w_a') \end{vmatrix}, \tag{17}$$

where only two electrons are involved in the transitions.

According to the Franck-Condon principle, an electronic transition in which the electronic excitation energy is lost to the lattice has small probability of occurrence, because of the condition of energy equality of the states ψ_1 and ψ_2. Under these conditions the transition probability P_{sa} should contain a Dirac delta function $\delta(E_1 - E_2)$, as a factor where $E_1 = w_s' - w_s$ and $E_2 = (w_a' - w_a)$ are the excitation energy on S in the state ψ_1, and that on A in the state ψ_2, respectively. The total probability of energy transfer from S to A becomes

$$P_{sa} = \left(\frac{2\pi}{\hbar}\right)\sum_1\sum_2 (g_s'g_a)^{-1} \times \int dw_a' \int dw_s \int dw_a p_a(w_a) \int dw_s' p_s'(w_s') \times |\langle H_1(w_s', w_a; w_s, w_a')\rangle_{12}|^2 \times \delta[(w_s' - w_s) - (w_a' - w_a)], \tag{18}$$

where g_s' and g_a are the degeneracies of the level of S and of the level of A, respectively.

Calculating by means of the S function and substituting $E = E_1 = E_2$, we have

$$P_{sa} = \left(\frac{2\pi}{\hbar}\right)\sum_1\sum_2 (g_s'g_a)^{-1} \times \int dE \int dw_a p_0(w_a) \int dw_s' p_s'(w_s') \times |\langle H_1(w_s', w_a; w_s' - E, w_{a+E})\rangle|^2. \tag{19}$$

The interaction between A and S can be expressed as the sum of all the Coulomb interactions of the outer electrons and the core of A with those of S, properly reduced by the dielectric constant K of the medium. Expansion of this sum in a Taylor series around the vector $\mathbf{R}$, which separates the nuclei of A and S, reduces the interaction Hamiltonian to the form

$$H_1(\bar{R}) = \left(\frac{e^2}{KR^3}\right)\{\bar{r}_s\cdot\bar{r}_a - 3(\bar{r}_s\cdot\bar{R})(\bar{r}_a\cdot\bar{R})/R^2\} + \left(\frac{3e^2}{2KR^4}\right)\{\sum_{i=1}^{3}(R_i/R)r_{ai}^2 r_{si}(-3+5R_i^2/R^2) + 10(XYZ/R^3)(x_a y_a z_s + x_a z_a y_s + y_a z_a x_s) + \sum_{i\neq j}^{3}\sum^{3}[(R_j/R) - 5R_i^2 R_j/R^3] \times [-r_{ai}^2 r_{sj} - 2r_{ai}r_{aj}r_{si}]\}, \tag{20}$$

where $\bar{r}_s = \sum_m \mathbf{r}_{s0m}$ refers to all electrons on S measured from its nucleus and so with $\mathbf{r}_a = \sum_n \mathbf{r}_{an}$.

Electric Dipole-Dipole Interaction

The first curly brackets in Eq. (20) are the dipole-dipole interaction term. Inserting this into Eq. (17), one can have

$$P_{sa}(dd) = \left(\frac{2\pi}{\hbar}\right)\sum_1\sum_2\left(\frac{e^4}{K^2R^6}\right)(g_s'g_a)^{-1} \times dE \int dw_a p_a(w_a) \int dw_s' p_s'(w_s') \times \langle\bar{r}_s\rangle\langle\bar{r}_a\rangle - 3(\langle\bar{r}_s\rangle\cdot R)|(\langle r_a\rangle R)|^2, \tag{21}$$

where the matrix element, $\langle\bar{r}_s\rangle$ stands for

$$\langle\bar{r}_s(w_s', w_s' - E)\rangle_{isfs}$$

and so with (r_a). Average of the absolute square of the matrix element in Eq. (21) over all possible orientations of $\mathbf{R}$ is $(2/3)|\langle r_s\rangle|^2|\langle r_a\rangle|^2$, where $|\langle r\rangle|^2 = |\langle x\rangle|^2 + |\langle y\rangle|^2 + |\langle z\rangle|^2$ so that

$$P_{sa}(dd) = \frac{4\pi e^4}{3hK^2R^6 g_s g_a}\sum_1\sum_2\int dE \times \left\{\int dw_s' p_s'(w_s')|\langle\bar{r}_s\rangle|^2\right\}. \tag{22}$$

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

By using the usual expressions for Einstein's A and B coefficients to evaluate the matrix elements contained here in terms of the experimentally measurable quantities such as oscillator strengths, absorption coefficients, and decay times, Dexter reduces Eq. (22) to the form,

$$P_{sa}(dd)=\frac{3h^4c^4Q_a}{4\pi R^6n^4\tau_s}\left(\frac{\mathcal{E}}{K^{\frac{1}{2}}\mathcal{E}_c}\right)^4\int\frac{f_s(E)F_a[E]}{E^4}, \quad (23)$$

with the absorption data for A and the emission data for S or to the form,

$$P_{sa}(dd)=\frac{3hc^2Q_aQ_s}{4\pi^3n^2R^6}\left(\frac{g_s}{g_s'}\right)\left(\frac{\mathcal{E}}{K^{\frac{1}{2}}\mathcal{E}_c}\right)^4\int\frac{f_s(E)f_a(E)}{E^2}dE, \quad (24)$$

which also contains the emission data for s, where Q_s, Q_a = areas under the absorption bands of S, A; τ_s = decay time for S emission; $\mathcal{E}$ = electric field at an isolated atom; $\mathcal{E}_c$ = electric field within the crystal; K = dielectric constant of the medium; $f_s(E)$ = shape of the observed S-emission band normalized so that $\int f_s(E)dE=1$; $F_a(E)=\sigma_a(E)/Q_a$; $\sigma_a(E)$ = absorption cross section of energy E by A, measurable at low concentrations of A, as the absorption coefficient per unit impurity (A) density in the medium.

This result is similar to Forster's[11] expression for the probability of (d-d) transfer between systems S and A at a distance R_{sa} apart,

$$P_{sa}=\frac{2c}{\pi^4n^2\bar{N}^2\bar{\nu}^2R_{sa}{}^6}\int f_s(\bar{\nu})\rho_a(\bar{\nu})d\bar{\nu}, \quad (25)$$

where $\bar{N}=6.06\times10^{23}$, $\bar{\nu}$ = wave number in cm^{-1}, $f_s(\bar{\nu})$ = molar emission intensity of S at $\bar{\nu}$, $\rho_a(\bar{\nu})$ = molar extinction coefficient of A at $\bar{\nu}$.

The transfer probability thus depends upon the strengths of the individual transitions as determined by decay times, absorption band areas Q, and the energy overlap of the emission and absorption bands of A and S. Neglecting nonradiative transfers in S and transfers from S to S, the following expression for the quantum yield for transfer is obtained,

$$\eta_T=p_{sa}\tau_s/(1+p_{sa}\tau_s). \quad (26)$$

By writing $\beta=(P_{sa}\tau_s)^{\frac{1}{2}}v$, where v is defined as $(4/3)\pi R^3$, one obtains $\eta_T=(\beta^2)/(v^2+\beta^2)$. Let C^+ be the number of density of lattice sites that can accommodate A and let v_0 be the volume excluded by the presence of the sensitizer. Writing $y=x_aC^+\beta$ one finds the average yield as

$$\bar{\eta}_T=y\exp(x_aC^+v_0)\{y^{-1}[\exp(-x_aC^+v_0)-1]+C_i(y)\sin y-S_i(y)\cos y+(\pi/2)\cos y\}, \quad (27)$$

where C_i and S_i are cosine and sine integral functions. For concentrations $x_a\lesssim1\%$ if $C^+=2.26\times10^{22}$ cm^{-3} ($y\lesssim20$); $\beta=8.26\times10^{-20}$ cm^{-3} the factors $\exp\pm(x_aC^+v_0)$ may be set equal to unity and one obtains

$$\bar{\eta}_{T'}=y(C_i(y)\sin y-S_i(y)\cos y+(\pi/2)\cos y\}. \quad (28)$$

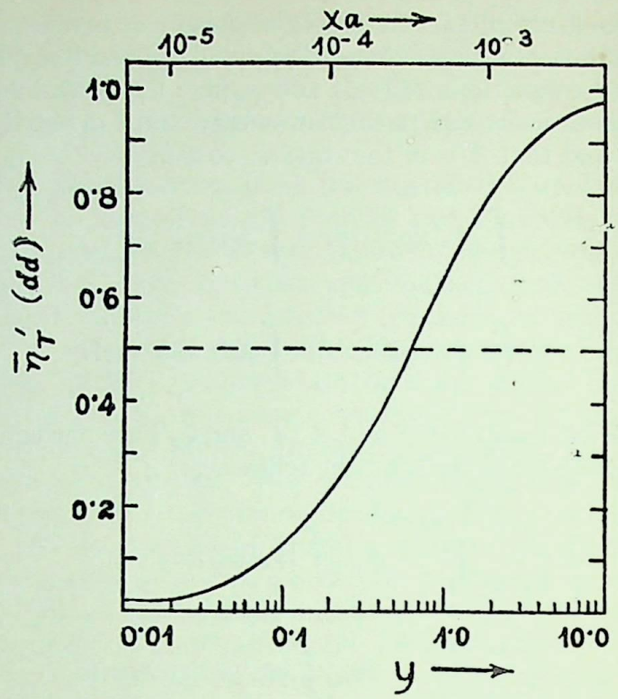


FIG. 1. Quantum yield for dd transfer as a function of the reduced concentration y of the activator. The upper abscissa represents the numerical values of the atomic concentration x_a in the typical case of NaCl crystals (after Dexter).

This function is shown in Fig. 1. For concentrations small enough that $y\ll1$, $\bar{\eta}_{T'}$ varies as $y(\pi/2+y\ln y-0.4228y)$ so that for small x_a, $\bar{\eta}_T$ increases linearly with concentration. For concentrations high enough to make $y\gg1$, $\bar{\eta}_{T'}\to1$.

The temperature dependence of the transfer probability can be understood from Eqs. (23) and (24). Generally Q_a, Q_s, or τ_s are not very dependent on temperature so that the main temperature dependence is determined by the temperature dependence of the energy overlap integral $\int f_s(E)F_a(E)/E^4$. Both f_s and F_a generally possess temperature broadening. In the event of perfect overlapping at all temperatures, the transfer probability decreases with increasing temperature. If, however, the absorption and emission bands have well separated centers, temperature broadening of the bands acts to produce greater overlap, and therefore grater transfer probability. At low temperatures the broadening being negligible as compared to natural zero-point widths, the transfer probability should not depend on temperature.

Electric Dipole-Quadrupole Interaction

When S makes an allowed dipole transition and A a prohibited quadrupole transition to calculate the transfer probability, one can insert in Eq. (19) the absolute square of the matrix element $\langle H_1\rangle$ computed from the second bracket of Eq. (20) and averaged over all

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

orientations of $\bar{R}$. Inserting

$$\langle|\langle H_1\rangle|^2\rangle_{Av}=\frac{9e^4\alpha}{4K^2R^8}|\langle r_s\rangle|^2|\langle N_a\rangle|^2 \quad \text{in Eq. (19),}$$

where $\alpha=1.266$ and $|\langle N_a\rangle|^2=|\langle x_a^2\rangle|^2+|\langle y_a\rangle|^2+|\langle z_a\rangle|^2+2|\langle xy\rangle|^2+2|\langle yz\rangle|^2+2|\langle zx\rangle|^2$ one obtains the following expression for the transfer probability:

$$P_{sa}(dq)=9e^4\pi\alpha/2\hbar K^2R^8g_s'g_a\sum_1\sum_2\int dE \times\left\{\int dw_s'p_s'(w_s')|\langle\bar{r}_s(w_s',w_s'-E)\rangle|^2\right\} \times\left\{\int dw_a p_a(w_a)|\langle N_a(w_a,w_a+E)\rangle|^2\right\}. \quad (29)$$

Relating, as before the bracketed quantities to either emission or absorption data when measurable or significant, for A or S, Dexter obtains

$$P_{sa}(dq)=\frac{135\pi\alpha\hbar^9c^8g_a'}{4\eta^6R^8\tau_s\tau_ag_a}\left(\frac{\mathcal{E}}{K^{\frac{1}{2}}\mathcal{E}_c}\right)^4\int\frac{f_s(E)F_a(E)}{E^8}dE. \quad (30)$$

If τ_s is not measurable or has no significance, one can use the expression

$$P_{sa}(dq)=\frac{135\alpha\hbar^6c^6Q_s}{4\pi n^4R^8\tau_a}\times\frac{g_a'g_s}{g_ag_s'}\left(\frac{\mathcal{E}}{K^{\frac{1}{2}}\mathcal{E}_c}\right)^4\int\frac{f_s(E)F_a(E)}{E^6}. \quad (31)$$

The ratio that $P_{sa}(dq)$ bears to $P_{sa}(dd)$ is calculated from Eqs. (30) and (23) to be $(45\alpha/4\pi^2)(n\lambda/R)^2\tau_a(d)/\tau_a(q)$. Since a quadrupole radiative transition has a probability of the order of $(a/\lambda)^2$ times a dipole radiative transition, one can put $\tau_a(d)/\tau_a(q)$ so that $P_{sa}(dq)$: $P_{sa}(dd)$ is of the order of $(a/R)^2$. Thus, resonance transfer can occur from a sensitizer to a nearby activator having a quadrupole transition in the proper frequency interval whereas direct absorption of radiation by the activator takes place with a probability only $\sim10^{-7}$ of that in an impurity with an allowed transition.

After defining γ by the relation $P_{sa}(dq)=\gamma^{8/3}v^{8/3}\tau_s$ and putting $y=\gamma C^+x_a$, one finds

$$\bar{\eta}_T=y\exp(x_aC^+v_0)\left\{y^{-1}[\exp(-x_aC^+v_0)-1]+\int_0^\alpha(e^{-vt}/1+t^{8/3})dt\right\} \quad (32)$$

and

$$\bar{\eta}_T'=y\int_0^\alpha(e^{-vt}/1+t^{8/3})dt. \quad (33)$$

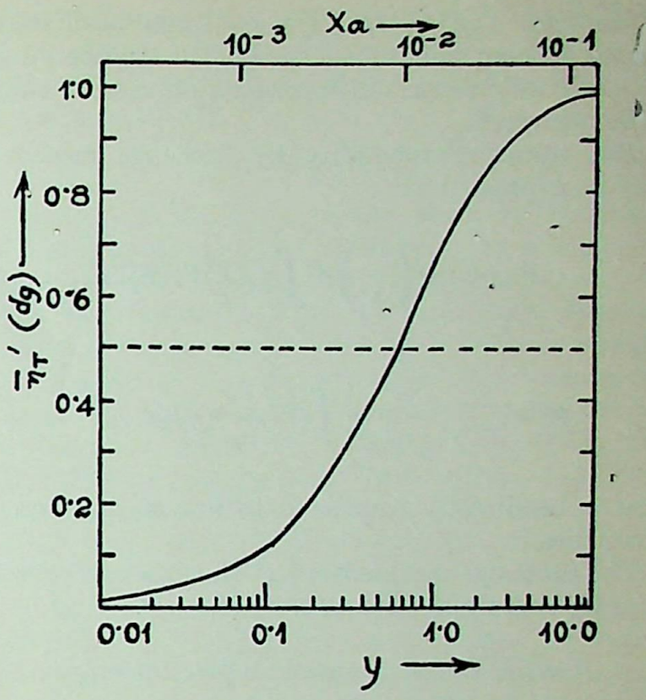


FIG. 2. Quantum yield for dq transfer as a function of the reduced concentration y of the activator, as in Fig. 1. The upper abscissa scale shows atomic concentrations in the typical case of NaCl crystals. The variation of $\bar{\eta}_T'$ with concentration is more rapid in the dq case than in the dd case (after Dexter).

This function is shown in Fig. 2. Since $P_{sa}(dq)$ depends on R more strongly than $P_{sa}(dd)$, $\bar{\eta}_T$ depends upon x_a more rapidly in the dq case than in the dd case.

The function $F_a(E)$ which is the absorption shape function of A cannot be experimentally determined because the absorption would not be observed.

Exchange Effects and Other Forbidden Transitions

Considering only the effects of electron spin, one can write out explicitly the matrix element of H_1 as $\psi(\bar{r},\bar{\sigma})=\varphi(\bar{r})\chi(\sigma)$ where $\chi(\sigma)$ are the spin wave functions.

$$\langle H_1\rangle=\int\varphi_s'^*(\bar{r}_1)\varphi_a^*(\bar{r}_2)H_1\varphi_s(\bar{r}_1)\varphi_a'(\bar{r}_2) \times\chi_s'^*(\bar{\sigma}_1)\chi_a^*(\bar{\sigma}_2)\chi_s(\bar{\sigma}_1)\chi_a'(\bar{\sigma}_2) -\int\varphi_s'^*(\bar{r}_1)\varphi_a^*(r_2)H_1\varphi_a'(\bar{r}_1)\varphi_s(\bar{r}_2) \times\chi_s^*(\bar{\sigma}_1)\chi_a^*(\bar{\sigma}_2)\chi_a'(\bar{\sigma}_1)\chi_s(\bar{\sigma}_2). \quad (34)$$

The first term in $\langle H_1\rangle$ is the Coulomb term treated before, and the second integral is an exchange integral with $H_1=e^2/Kr_{12}$. If the X's are left for the moment the integral can be understood as representing the electrostatic interaction between the two charge clouds

$$Q'(r_1)=\varphi_s'^*(\bar{r}_1)\varphi_a'(\bar{r}_1) \quad \text{and} \quad Q(\bar{r}_2)=\varphi_a^*(\bar{r}_2)\varphi_s(\bar{r}_2).$$

Each function φ dies off exponentially with distance from S or A and each product Q will not be sizeable throughout all space except when r_{12} is small so that

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

the integral $\int Q'(\bar{r}_1)(1/r_{12})Q(\bar{r}_2)d\tau_{12}$ may be of sizeable value (R being small) even though the overlap integral $\int Q(\mathbf{r})d\tau$ which enters in normalization, is small enough to be neglected.

The transfer probability by exchange mechanism may be expressed as

$$P_{sa}(ex)=\left(\frac{2\pi}{h}\right)Z^2\int f_s(E)F_a(E)dE, \qquad (35)$$

where

$$Z^2=\sum_1\sum_2\frac{e^4}{g_s{}'g_a\kappa^2}\left|\int Q'(r)\frac{1}{r_{12}}Q(\bar{r}_2)d\tau_{12}\right|^2_{1.2}$$

cannot be directly connected to optical experimental quantities.

The exchange mechanism has an efficiency in energy transfer slightly less than the dq transfer.

Electric Dipole-Magnetic Dipole Interaction

The transfer mechanism arising from an electric dipole-magnetic dipole interaction may be treated by including magnetic fields in the Hamiltonian. The most important in a so-constructed H_1 is $e\hbar/2mc\bar{H}\cdot(\bar{r}_a\times\bar{\nabla}_a+i\bar{\sigma}_a)$, where H is the magnetic intensity at A produced by motion of charge at S. The largest term in H being of the order $(e/c)\bar{R}\times\bar{V}_s/R^3$ or $(e\hbar/mc)\bar{R}\times\bar{\nabla}_s/R^3$. Thus $\langle H_1\rangle$ has the magnitude $(e^2\hbar^2/2m^2c^2R^2)\langle\bar{\nabla}_s\rangle$ or because $\langle\bar{\nabla}_s\rangle=mE\langle\bar{r}_s\rangle/\hbar^2$ one finds

$$|\langle H_1(em)\rangle|^2/|\langle H_1(dd)\rangle|^2\sim(e^2/hc)^4E^2/(e^2/R)^2. \qquad (36)$$

Thus for the nearest neighbors $P_{sa}(em)$ is of the order of 10^{-8} times $P_{sa}(dd)$, small enough to be neglected.

Host Sensitization

Most of the discussion of impurity sensitization applies also to host sensitization, i.e., sensitization of an activator by the lattice itself in which it is embedded. In this case X_s, being equal to unity, is not small enough to forbid S-S transfer. When S has an allowed transition the transfer mechanism from S to S would be a (d-d) process, and depending on the degree of forbiddenness of transitions in A, either a dd process, a competition between dq and exchange or exchange from S to A.

Transference of excitation energy by dipole-dipole interaction should be carefully distinguished from a cascade phenomenon. In the latter process where a photon emitted by S is reabsorbed by A, the quantum yield for transfer depends not only on the concentration (A) but also on the size and shape of the system, as seen from (16).

The quantum yield for transfer by the cascade mechanism can be expressed as follows in terms of the crystal dimension,

$$\bar{\eta}_T(C)=Av\int f_s(E),$$

where A, v, $f_s(E)$, C, and (E) are as defined before and the average is taken over the linear dimension l of the sample. $1-\eta_{0T}$ is the probability that a photon emitted by S will escape from the system and is then equal to η_{cs}. The quantity $C(A)l$ represents the number of impurity molecules A per cm^2 projected on a plane perpendicular to the path of the photon. If we compare a spherical system containing a concentration (A) of A with the same system spread out in a thin film as regards quantum yield, it will be greater in the former case than in the latter, if the transfer takes place by the cascade process whereas it will be independent of the shape and size of the sample if the transfer takes place by dipole-dipole interaction.

(c) Exciton Migration

Exciton or excitation wave, a concept first introduced by Frenkel[13] to mean quick transfer of excitation energy from one constituent to another of a crystal lattice, is a product of the mathematical problem of calculating the excited states of a lattice of interacting atoms or molecules. The physical model used by Frenkel, Peierls,[14] and subsequently, by Slater and Shockley[15] was simple. It consists of a system of identical atoms each containing one electron arranged in simple crystalline array. The electron spin is neglected and both the lowest and the first excited states are assumed to possess no degeneracy. Let atoms be located at the points defined by

$$\mathbf{r}(n)=n_1\tau_1+n_2\tau_2+n_3\tau_3, \qquad (37)$$

where the τ are the primitive translations of the lattice and the n range over all integral values. Let ψ_1 and ψ_2 denote the normal and the excited wave functions of the particular electron at $\mathbf{r}(n)$ and assume that there is so little overlapping of the wave functions on different atoms that the ψ are practically the same as the atomic wave functions. Consequently, ψ_n and $\psi_n{}'$ are orthogonal to each other and to the wave functions of electrons on other atoms. A wave function for the lowest state of the entire crystalline array, which may be expressed as a determinant of the form

$$\psi_0=\frac{1}{(N!)^{\frac{1}{2}}}\begin{vmatrix}\psi_1(\bar{r}_1)&\cdots\psi_1(\bar{r}_N)\\ \psi_N(\bar{r}_1)\cdots\psi_N(\bar{r}_N),\end{vmatrix} \qquad (38)$$

where N is the total number of electrons and atoms. The mean energy of this state is $E_0=\int\psi_0{}^*H\psi_0d\tau$, where H is the Hamiltonian of the entire system.

Let us now consider the wave function ψ_f of the entire system for the state in which one atom is excited from

[13] J. Frenkel, Phys. Rev. **37**, 1276 (1931); Physik. Z. Sowjetunion, **9**, 158 (1936).
[14] R. Peierls, Ann. Physik **13**, 905 (1932).
[15] J. C. Slater and W. Shockley, Phys. Rev. **50**, 705 (1936).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

the state ψ_f to the state ψ_f'

$$\psi_f=\frac{1}{(N!)^{\frac{1}{2}}}\begin{vmatrix}\psi_1(\bar{r}_1)\psi_1(\bar{r}_2) & \cdots\psi_1(\bar{r}_N)\\ \psi_2(\bar{r}_1)\psi_2(\bar{r}_2) & \cdots\psi_2(\bar{r}_N)\\ \cdots\cdots\cdots\cdots & \\ \psi_f(\bar{r}_1)\psi_f(\bar{r}_2) & \cdots\psi_f(\bar{r}_N)\\ \psi_N(\bar{r}_1)\psi_N(\bar{r}_2)\cdots\psi_N(\bar{r}_N). & \end{vmatrix} \tag{39}$$

We cannot regard this system of wave functions as proper representations of the excited states of the crystal because they do not make

$$E_{gf}=\int\psi_g^*H\psi_f d\tau \tag{40}$$

vanish when $m\neq n$.

The N-dimensional matrix constructed from the E_{gf} can be diagonalized by finding those linear combinations ψ' of the ψ_f which are of the form

$$\psi'=\sum a_f\psi_f, \tag{41}$$

where the a_f satisfy the following equation

$$\sum_g a_g E_{gf}=E_{af}'. \tag{42}$$

Since crystal symmetry makes E_{gf} dependent only upon the difference between g and the f values, the N equation (42) can be reduced to the same form if we substitute

$$a_g=a_k e^{2\pi i\mathbf{k}\cdot\mathbf{r}(g)} \tag{43}$$

where $\mathbf{K}$ is a vector in the reciprocal lattice of the crystal. This reduces (42) to

$$E_K'=E_{ff}+\sum E_{f,f+s}e^{2\pi i\mathbf{k}\cdot\mathbf{r}(s)} \tag{44}$$

where $E_{f,f+s}=E_{g+m,g+m+s}$ where m, s are arbitrary integer sets.

After substituting Eq. (44) in (41), we have the normalized wave function

$$\psi_K=\frac{1}{N^{\frac{1}{2}}}\sum e^{2\pi i\mathbf{k}\cdot\mathbf{r}(f)}\psi_f. \tag{45}$$

A wave function of this type represents the fact that the locality where the atom is excited, moves about the lattice with a wave number $\mathbf{k}$. In other words, the excited state is not localized on the particular atom that is excited, but is handed over from one constituent of the lattice to its neighbors.

By virtue of the assumption of weak overlapping we assume that E_{gf} vanishes for all but the nearest neighbors. If E_{gf} has the value ϵ in this case, (44) reduces to

$$E_K'=E_{ff}+\epsilon\sum e^{2\pi i\mathbf{k}\cdot\mathbf{r}}, \tag{46}$$

where $\mathbf{r}$ is to range over the nearest neighbors.

For a cubic lattice with lattice constant a, Eq. (46) takes the simple form

$$E_K'=E_{ff}+\epsilon(\cos 2\pi k_x a+\cos 2\pi k_y a+\cos 2\pi k_z a). \tag{47}$$

This means that the excited levels constitute a band having a spread of the order of magnitude ϵ.

It can be shown that the wave of excitation moves with a group velocity,

$$V=(1/h)\,\mathrm{grad}_K E, \tag{48}$$

and also that it carries no current. We may thus look upon the excitation wave as if it were an uncharged particle created by exciting the crystal and capable of moving about the lattice. This concept is due to Frenkel who called the wave of excitation an exciton.

It is instructive to look into Wannier's[16] model of treating the excitation bands in insulators because of its simplicity. If an electron is removed from the highest filled band of an insulator a positive charge, or positive hole, results. This positive hole may be filled up by an electron from an adjacent lattice constituent, and thus the positive hole may be able to roam about freely in an undistorted lattice. Though the Bloch picture endows the excited electron with freedom of movement independently of the positive hole, one should attract the other with a force that is Coulomb-like at large distances. It has been shown by Wannier that at the discrete levels of the hydrogen atom the excitation bands arise because of the motion of the electron and hole, one about another in closed states, and the different levels in a band behave as if they were different transitional levels of an excited hydrogen atom. This picture allowed Wannier to obtain a set of simple equations from which to derive the wave functions and energy levels of an exciton.

The migration of excitation energy by the exciton method should be carefully distinguished from that by dipole-dipole interaction. When excitons are generated in a crystal, the excitation energy is propagated through the lattice with a given wave-number vector $\mathbf{K}$ and the total wave function of the crystal describes a moving state of excitation energy of previously defined momentum, but indeterminate coordinates. The interaction Hamiltonian which describes the moving state of excitation energy is in form the same as that which determines the probability of transfer in a quantum mechanical collision process. The fundamental difference between them is that the former process is the result of interaction between an excited atom or molecule and all other like units, while the latter process results from interaction between one excited sensitizer and a normal activator. The latter is thus a one-step process which may take place over a distance of many angstroms in a neutral solvent, while the former may be regarded as a stepwise propagation of excitation energy from molecule to molecule covering many successive independent acts and terminates by one of a number of possible events. The motion of an exciton may be terminated by trapping at an impurity, by decay with characteristic emission of the solvent, or by

[16] G. H. Wannier, Phys. Rev. 52, 191 (1937).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

a radiationless transition. The transfer of energy to an impurity by exciton migration obeys a concentration dependence that is entirely different from that in case of the single step process of collision. The theory of exciton propagation in an insulating crystal has been extended by Heller and Marcus[17] to include cases where appreciable overlapping of the electronic wave functions of nearby crystal constituents is absent. Whereas in the primary works of Frenkel, and later Slater and Shockley the ease of exciton propagation was ascribed to the overlap of the excited state wave functions of the neighboring crystal constituents, Heller and Marcus have shown that very good propagation may result, notwithstanding poor overlap, from the coupling produced by the dipole-dipole terms of the interaction Hamiltonian which give rise to the van der Waals interactions in second-order perturbation theory. When one of the lattice constituents is excited to such a state that transition between it and the ground state is associated with a large oscillator strength, the strong coupling between this excited constituent and its identical neighbors in the ground state allows the migration of energy through lattice. This happens because the translational symmetry of the lattice permits of solution of the corresponding Schrödinger equation in the form of moving waves of excitation, analogous to Frenkel's exciton waves. One of their important results is that even in an idealized cubic crystal there exists anisotropy in the propagation of excitons. Two interesting cases of exciton propagation are distinguished, those in which the induced dipole moment $\mathfrak{u}$ is parallel and perpendicular, respectively, to the vector $\mathbf{k}$ of the exciton wave (Fig. 3). Provided there is no interference from polar modes of lattice vibration, exciton waves of the type $\mathfrak{u}\cdot\mathbf{k}=\pi/2$ are excited by absorption of light waves, since in that case the electric vector and the propagation vector for light quanta are normal to each other. Conversion of excitons from the type $\mathfrak{u}\cdot\mathbf{k}=\pi/2$ into the type $\mathfrak{u}\cdot\mathbf{k}=0$ and vice versa may result from scattering collisions with phonons or other imperfections in the lattice. It is of interest to note that the excitons for which $\mathfrak{u}\cdot\mathbf{k}=\pi/2$ have an energy increasing with k for all prominent directions of propagations, namely (100), (110), and (111) in a simple cubic lattice, and thus these excitons created by light waves behave as if they possess a positive mass. Those excitons which are polarized parallel to $\mathbf{k}$ behave in the reverse way and as if they possess a negative mass. The magnitude m' of the effective mass of an exciton is given by

$$m'\simeq mR_0/fn_0a_e,$$

where m is the electronic mass, f is the oscillator strength of the transition between the ground state and the excited state, a_e is the equivalent Bohr radius for the excited atom, and $R_0=(3\pi/4n_0)^{\frac{1}{3}}$, n_0 being the atomic density. Thus, the larger the oscillator strength the quicker is the propagation of radiationless migration of excitation energy from one atom or complex to another atom or complex, attributed by Cario, Franck, and Forser to dipole coupling. The speed of migration varies inversely as the sixth power of their separation. In the case of exciton propagation, by virtue of the very high density of states, the stationary state cannot be located on one particular excited atom at a time and so the appropriate model is much like the electron wave packet in metals, in which collisions changing $\mathfrak{u}$ and $\mathbf{k}$ will take place during the lifetime of the excitations. Following this approach one can think of a typical maximum diffusion distance of an exciton wave packet d as given by

$$d\leq(\langle v^2_{\text{thermal}}\rangle\tau_{\text{thermal}}\tau_{\text{optical}})^{\frac{1}{2}}$$
$$\approx(10^{14}\times10^{-14}\times10^{-8})^{\frac{1}{2}}=10^{-4}\text{ cm.}$$

The total length of the path traversed may be of the order of centimeters. The exciton may be deprived of its energy long before the optical lifetime has expired by a competing process.

All these conclusions are reached on the basis of the assumption of a small amount of overlapping of the excited wave functions of the electrons. In certain ionic and molecular crystals in which the electron wave functions are strong at the widely spaced lattice sites and weak in between, this assumption is an appropriate qualitative description of the state of affairs.

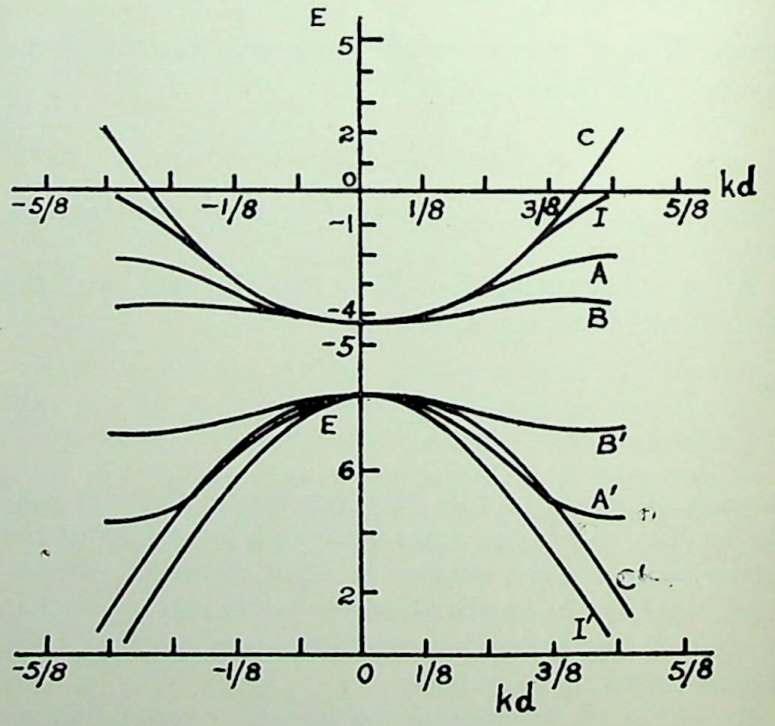


FIG. 3. The energy of exciton waves as a funcion of wave number (after Heller and Marcus). The upper curves refer to excitons for which the electric dipole moment is perpendicular to the direction of propagation, whereas the lower family refers to excitons characterized by polarization parallel to the wave number vector k, d is $(2n_0)^{-\frac{1}{3}}$ where n_0 is the density of ions contributing to the exciton propagation. The energy unit is $\mu_0^2n_0$ where μ_0 is the electric dipole moment of the excited state. Curves A, B, C in the upper diagram and A', B', C' in the lower are for k vectors in the (100), (110), and (111) directions, respectively. Integral approximations in which the excitons behave isotropically are shown in curves I and I'.

[17] W. R. Heller and A. Marcus, Phys. Rev. 84, 809 (1951).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

(d) Electron and Hole Migration

Production of electrons and holes and their migration were first suggested by Mott and Gurney[18] to account for electronic behavior in insulating crystals. Such a crystal is characterized by a full band of energy levels, named the valence band, separated by a finite energy gap W from an empty conduction band (Fig. 4). If absorption of a quantum of radiation raises an electron to the conduction band, the resulting vacancy in the valence band, called a positive hole, may be filled up by an electron from another constituent whose full band lies above the starting band of the absorption processes in Fig. 5(a) and thus may be caught in a trap. When the free electron from the conduction band finally recombines with its holes trapped in the new host, there results a new emission. It may also happen that the excited electron temporarily falls into an intermediate vacant band below the conduction band as in Fig. 5(b) and the holes become mobile through the lattice until it recombines with its trapped counterpart with the emission of a new band. The energy set free by the trapping of the electron or its hole may be emitted as infrared radiation or dissipated as heat. The traps of the electrons or holes are localized—the energy levels being produced by an activator or lattice flaws, or cracks. Such crystals exhibit photoconductivity because of the mobility of an electron in the conduction band or because of its positive holes in the valence band.

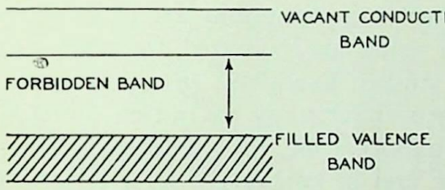


FIG. 4. Schematic representation of the filled and unfilled levels.

When energy transfer takes place over a long distance, each of these processes requires a transport mechanism:

(1) In a quantum mechanical collision process, the collision partners themselves may become vehicles of long distance transportation of energy.

(2) The excitation energy may travel through the lattice by the migration of an excited electron and its holes freed from each other.

(3) In the exciton mechanism, the transport of energy through the lattice takes place in the shape of excitation wave.

(4) Photon absorption is not a localized process but may occur over larger distances.

Although the same mechanism may account for energy transport and its transfer such as in CaWo—Sm (mechanism 3) and the Zn—S luminiphors (mechanism 2), this is not indispensable. It is not improbable that

[18] N. F. Mott and R. W. Gurney, *Electronic Process in Ionic Crystals* (Oxford University Press, New York, 1949).

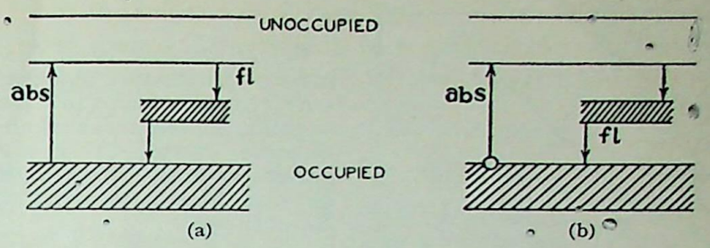


FIG. 5. The excited electron falls into an intermediate vacant band below the conduction band.

excitation energy is transported by say migration of electrons and holes, whereas the final transfer is effected by, say, a collision process.

2. FLUORESCENCE TRANSFER IN ORGANIC MIXED CRYSTALS

In 1949 Franck and Livingston[9] summarized the facts on fluorescence in organic mixed crystals brilliantly. Since then many more publications have appeared and considerable new results are now at hand, so it seems desirable to discuss the subject afresh. The present discussion is confined to organic crystals in general and naphthacene in anthracene in particular for the following reasons: (a) a considerable amount of experimental results and theoretical discussion have been published on it which have made it a matter of controversy; (b) we are studying the matter in this laboratory and some interesting quantitative results have been obtained by us and our collaborators.

(a) Exciton and Absorption

Interest in transfer of energy from one molecule to another in an aromatic solid seems to grow continuously from the anthracene-naphthacene problem. Many workers[19–22] have observed that in the pure crystalline state naphthacene hardly fluoresces, whereas dissolved in solid anthracene or chrysene, it fluoresces brilliantly with its characteristic yellow-green color and the quenching of host crystal fluorescence. The quantum yield of its photofluorescence increases from 0.04 in a pure crystal to about 1 in anthracene.[23] Bowen[20] suggested that the mechanism of transfer of excitation energy from anthracene to naphthacene molecules embedded in the anthracene lattice possesses a close resemblance to that observed in a typical inorganic phosphor, and that the energy of excitation travels by the exciton mass. On the other hand, Ganguly's[24] investigations into the absorption and fluorescence phenomena of these aromatic hydrocarbons under different physical conditions, particularly the depend-

[19] Dufraisse and Horclois, Bull. soc. chim. 1888 (1936).
[20] E. J. Bowen, Nature 142, 1081 (1938).
[21] S. C. Ganguly, Nature 151, 3841 (1943).
[22] S. C. Ganguly, Nature 153, 652 (1944).
[23] E. J. Bowen and A. H. Williams, Trans. Faraday Soc. 35, 765 (1939).
[24] S. C. Ganguly, J. Chem. Phys. 13, 128 (1945); Indian J. Phys. 19, 22B (1945). Doctoral thesis, Calcutta University, 1949.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

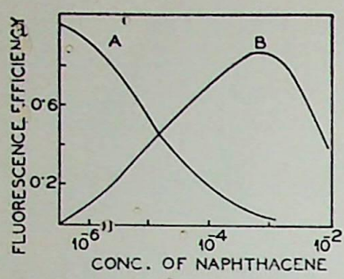


FIG. 6. Dependence of fluorescence efficiency of naphthacene concentration in the naphthacene contaminated anthracene crystal. *A* is for anthracene fluorescence, *B* is for naphthacene fluoroescence (after Bowen).

ence of the latter upon the wavelength of exciting radiation suggests that the exciton mechanism is unnecessary in this case and that the enhanced fluorescence of naphthacene could be related to the excitation of the naphthacene molecules by way of absorption of the exciting radiations as well as the fluorescent light of anthracene. This suggestion was based on the following observations:

(i) When excited by light of wavelength from 4030 to 4350 A, green anthracene crystal or a green chrysene crystal emits the characteristic naphthacene fluorescence, except that it is shifted toward the red due to the solvent effect. Naphthacene molecules absorb these exciting wavelengths but the anthracene lattice does not.

(ii) Since the absorption region of naphthacene overlaps the emission region of anthracene (3650 to 5000 A), the naphthacene molecules absorb not merely the exciting radiation 3650, but must also absorb the fluorescence radiations of anthracene.[22]

(iii) There is positive evidence for photon absorption by naphthacene molecules followed by photon emission, and lack of evidence for the exciton mechanism.

Almost simultaneously Bowen and his co-workers[25–28] were investigating the effect of naphthacene concentrations, particle size etc., on the intensities of anthracene and naphthacene fluorescence. They observed the following.

(i) At a concentration of 10^{-5} naphthacene molecule per anthracene molecule, the green naphthacene fluorescence is visible and at 10^{-4} naphthacene molecules per anthracene molecule it grew strong. Lack of molecular proximity and orientation in liquid solution inhibited the transfer so that a green crystal of anthracene with a low concentration of naphthacene when dissolved in benzene did not give the characteristic green fluorescence of naphthacene under the same excitation.

(ii) The curves (Fig. 6) relating the intensities of anthracene and naphthacene fluorescence to the concentration of naphthacene in an anthracene crystal could be represented by $n=1/1+kx$ and $1-\eta=kx/1+kx$, where n=quantum efficiency of fluorescence of host lattice; $k=a$ constant for a pair of compounds in the host and guest relation, excited by a particular radiation; and x=guest concentration in g per g of the host. For solid solutions of anthracene in naphthalene k was found to be 5×10^4 when excited by 2900 to 3200 A radiations at 20°C, for naphthacene in anthracene crystal excited by 2900 to 3200 A, k was found to be 2×10^5 at 20°C.

(iii) The curves of intensities of fluorescence *vs* naphthacene concentrations in anthracene crystal do not differ greatly from those in solutions precipitated from one solution. When separate mixtures of the two are taken, there is produced a great difference.

(iv) The intensity ratio of the 0—1 and 0—2 bands in the fluorescence of anthracene remains unaltered for all concentrations of naphthacene in anthracene crystals.

(v) The intensity of the 0—1 band compared to that of the 0—2 band in the fluorescence of anthracene increases as the particle size is diminished and finally approaches that for dilute solution in benzene.

These observations were regarded by Bowen and his co-workers as evidence supporting the exciton mechanism in these crystals. Later it was shown by Birks[7] that the observations (i) and (ii) could not lead to this unambiguous conclusion.

Lipsett and Dekker's[29] investigtions are of interest in that, apart from using an improved technique of measurement, they studied band-wise the spectral intensity of fluorescence under x-ray excitation emitted by solid solutions of naphthacene, 1,2,5,6-dibenzanthracene in naphthalene and in anthracene as a function of concentration. The relative band intensity of impurity fluorescence increased with the rise of concentrations in the host lattice, reached its maximum, and then fell off in a manner shown in Fig. 7, for 1,2,5,6-dibenzathracene in naphthalene. Considering the high efficiencies of energy transfer at concentrations as low as 1.01×10^{-4} mole naphthacene per mole naphthalene and 2.93×10^{-4} mole naphthacene per mole anthracene

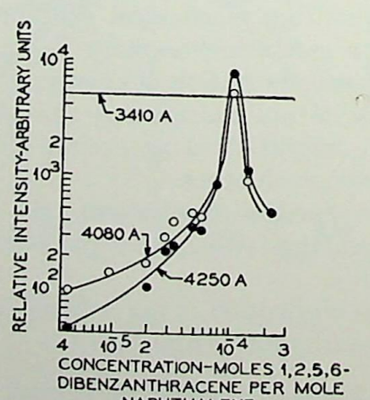


FIG. 7. Intensity of the 4080 and 4250 A bands of 1,2,5,6-dibenzanthracene fluorescence relative to the 3410 naphthalene fluorescence band as function of concentration (after Lipsett and Dekker).

[25] E. J. Bowen, Nature **153**, 663 (1944).
[26] E. J. Bowen and E. Mikiewicz, Nature **159**, 706 (1947).
[27] E. J. Bowen and P. D. Lawley, Nature **164**, 572 (1949).
[28] Bowen, Mikiewicz, and Smith, Proc. Phys. Soc. (London) **A62**, 26 (1949).
[29] F. R. Lipsett and A. J. Dekker, Nature **173**, 736 (1954).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

for maximum transfer, they discard the absorption and re-emission as unlikely and regarded the exciton mechanism as more probable. Subsequently, they observed that the intensity of fluorescence under x-ray excitation of anthracene molecules embedded in the naphthalene lattice increased steadily with concentration even after the naphthalene fluorescence was quenched. They attributed this complex behavior to a vigorous fluorescence of anthracene molecules by themselves in addition to the transferred excitation energy from naphthalene molecules to anthracene molecules by way of exciton. The anthracene bands at 3505 and 3700 A arose primarily from the transferred excitation energy and those at 3895, 4200 A, etc., from a combination of transferred excitation energy and molecular fluorescence.

Recent investigations of fluorescence of solid solutions of nine aromatic hydrocarbons in anthracene and pyrene by Northrop and Simpson[30] have supplied significant quantitative information as regards the transfer of excitation energy from solvent to solute molecule. The degree of energy transfer is discussed in terms of two experimentally determined quantities, namely, the quenching factor and the transfer efficiency, defined by

$$Q=(S-S')/S \quad \text{and} \quad T=I/(S-S'),$$

where S = integrated emission of pure solvent molecules, S' = integrated emission of a pure solvent containing solute molecules, and I = integrated emission of solute molecules. These investigators found (Fig. 8) that for anthanthrene dissolved in anthracene the quenching factor is proportional to concentration up to 10^{-2} M and at higher concentration becomes less than what is expected from the proportionality relation. The behavior of transfer efficiency with increasing concentration C is given in Table II. Thus all the quanta quenched in anthracene fluorescence do not reappear as anthranthrene fluorescence. There is an optimum concentration, 10^{-3} M in this case, at which T is maximum, unity in this case. For molecules of smaller size studied by them such as naphthacene, pentacene, perylene, etc., T was greater at 10^{-3} M than at 10^{-4} M. Larger molecules such as iso-violanthrene and ovalene display a reversed tendency. The concentration dependence of Q is semi-quantitatively explained by Northrop and Simpson on the basis of exciton migration. When the matrix lattice of the solvent is not contaminated with impurity molecules a solvent molecule may return to the ground state by emitting fluorescence, by nonradiative transitions causing internal quenching, or by transferring its excitation energy to a like neighbor and let α, β, and γ be the probabilities per unit time of the individual processes, respectively. Considering that an exciton continues to move through the crystal until it ends its life it is potentially able to cause fluorescence, they obtained the total probability of emission of one fluorescence quantum per one molecule initially excited as $S=\alpha/(\alpha+\beta)$ where the matrix lattice contains foreign molecules of concentration c, there is a probability, q, say, that the exciton may be captured during its propagation by an adjacent foreign molecule. This reduces the probability of emission by a solvent molecule to

$$S'=\alpha/\alpha+\beta+\gamma cq,$$

where s and s' are proportional to S and S'. Therefore,

$$Q=\frac{S-S'}{S}=\frac{s-s'}{s}=\frac{\gamma cq}{\alpha+\beta},$$

which predicts that Q is proportional to c in agreement with the experimental results.

The investigations of Northrop and Simpson point out that there is no correlation between quenching and the overlap between solvent emission and impurity emission and impurity absorption. There is very little overlap in the case of iso-violanthrene embedded in the anthracene lattice and yet iso-violanthrene quenches anthracene fluorescence strongly with $Q=11$ and 31 at $C=10^{-4}$ and 10^{-3} M, respectively. Naphthacene has a very strong absorption overlapping the emission of pyrene but there was observed no quenching or transfer when at high concentration. These facts are of great importance as they rule out the possibility of energy transfer by collision processes. The probability of energy by collision is proportional to $\int \nu^{-4}E(\nu)A(\nu)d\nu$, where $E(\nu)$, $A(\nu)$ are the fluorescence spectrum of the excited molecule and the absorption spectrum of the acceptor molecule.

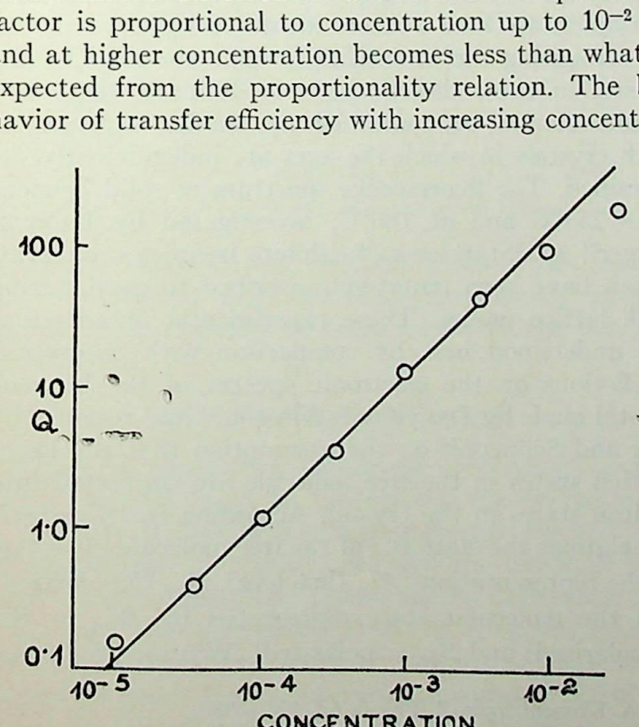


FIG. 8. Dependence of the quenching factor Q on the concentration of impurity (after Northrop and Simpson).

TABLE II.

C	0	10^{-5}	3.3×10^{-5}	10^{-4}	3.3×10^{-4}	10^{-3}	3.3×10^{-3}	10^{-2}	3.3×10^{-2}	10^{-1}
T	0	0.57	0.75	0.81	0.95	1.02	0.93	0.75	0.50	01.4

[30] D. C. Northrop and O. Simpson, Proc. Roy. Soc. (London) A234, 136 (1956).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

3. ELECTRONIC STATES OF AROMATIC CRYSTALS

One of the glaring features of early investigations of energy transfer is that the question whether transference of excitation energy by way of the exciton in aromatic crystals is feasible in view of weak intermolecular binding in them, was not attacked by a consideration of the nature of electronic states in these crystals. In spite of the works of Heller and Marcus,[17] some[7,9] had the idea that exciton propagation required strong intermolecular coupling. As there existed at that time little evidence of such intermolecular coupling in these crystals, often the possibility of exciton propagation was ruled out.

Theoretical works on the electronic states of molecular crystals, begun not long ago, are all based on the elementary theory of excited states of a lattice of interacting atoms or molecules given by Frenkel,[13] later by Slater and Shockley[15]; and improved by Peierls[14] to include the effect of lattice vibrations. One of the principal results of this theory is that in aromatic hydrocarbon crystals beneath conduction bands there occur nonconducting excited states called exciton bands. Just as the conduction bands are analogous to ionization levels of atoms or molecules, these excitation bands in a crystals are analogous to the excited states of free atoms or molecules and are characterized by the fact that the optical excitation to these bands is accompanied by propagation of this state of excitation from one lattice constituent to another, by what is called an "exciton."

Nonconducting excited states or exciton states as they are generally called, arise in two ways. In weakly bound solids (molecular crystals such as anthracene, etc.), the exciton bands correspond closely to the lowest excited states of the constituent atoms or molecules in the free state. Secondly, the strongly bound ionic or covalent crystals, where the constituents have lost their individuality, arising in the ionic case from the removal of an electron from one atom to a neighbor, or by the sharing of electrons with neighbors in the covalent case. An application of Frenkel's exciton theory to the calculation of energy states of molecular crystals was made first by Davydov.[31] He envisaged a weak coupling model of a molecular crystal characterized by an intermolecular interaction energy small compared to the intramolecular energy so that crystallization does not appreciably affect the electronic structure of a molecule. This is required by the fact that the broad features of the visible and ultraviolet spectra are but little changed by the process of crystallization. One of the most significant results reached by his calculations is that from each excited level of a free molecule there originates a band of energy levels in the crystal. And what is more important, corresponding to a transition in the free molecule there may take place several optical transitions in the crystal, notwithstanding the nondegeneracy of the free molecular levels. This results in multiplicity of lines corresponding to one optical transition in the free molecule. This effect, called Davydov splitting, must be carefully distinguished from the two other well-known effects of crystal field. One of these is the splitting of levels that are degenerate in the free molecule by the crystal field. This effect has been discussed by Bethe. It has found wide application in the interpretation of the spectra of rare earth crystals. The other is the occurrence of crystal-field-induced transitions which are altogether forbidden in the free molecules.

We now discuss the extent to which the exciton effects occur in the crystals of aromatic hydrocarbons in accordance with theoretical predictions.

(a) Benzene Crystal

The benzene crystal being the simplest of the aromatic crystals, its importance in interpreting the nature of the solid state of the aromatic hydrocarbons is unique. The electronic spectra have been investigated both experimentally and theoretically. The first transition in solid benzene was first observed by Kronenberger.[32] The spectrum exhibited some degree of fine structure not accountable without assuming some crystal effects, either Davydov splitting, or combination with lattice modes. Recently Broude *et al.*[33] have investigated the absorption spectrum of single crystals of benzene using polarized light and detected two perpendicularly polarized lines, separated by 25 cm^{-1} and corresponding to the lowest excited singlet state of the free molecule. Inasmuch as these workers identified the crystal axes by comparing the intensities of light absorbed by the crystal in these directions, it is desirable that their results should be confirmed by similar experiments with crystals in which the axes are independently determined. The fluorescence spectrum of solid benzene at −253°C and at 180°C, investigated by Kronenberger[32] and Starkiewicz[34] exhibits frequency intervals which have been tentatively ascribed to combination with lattice modes. These experimental observations are understood best by comparison with theoretical predictions on the electronic spectra of the benzene crystal made by Davydov,[31] Winston,[35] and recently by Fox and Schnepp[36] on the assumption that the lower excited states in the free molecule are converted into exciton states in the crystal. According to Davydov's calculations the state B_{1u} of the free molecule gives rise to the representations A_{1u} (inactive), B_{2u} (b polarized) and the molecular state represented by B_{2u} to B_{1u} (a polarized) and B_{3u} (c polarized). Winston confirmed

[31] S. S. Davydov, J. Exptl. Theoret. Phys. (U.S.S.R.) **18**, 210 (1948); **21**, 673 (1951).

[32] A. Kronenberger, Z. Physik **63**, 494 (1930).

[33] Broude, Medvedev, and Prikhotjko, J. Exptl. Theoret. Phys. (U.S.S.R.), **21**, 665 (1951).

[34] J. Starkiewicz, Compt. rend. soc. polon. phys. 4, 201, (1929).

[35] H. Winston, J. Chem. Phys. **19**, 156 (1951).

[36] D. Fox and O. Schnepp, Phys. Rev. **96**, 1196 (1954); J. Chem. Phys. **23**, 76, 7 (1955).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

these results and also found that the forbidden 0—0 transition of the free molecule is not prohibited in the crystal. Fox and Schnepp calculated the exciton band energies and band widths in the benzene crystal corresponding to the B_{1u} and the B_{2u} states of the free molecule, to which dipole transitions from the ground state are allowed. They used the x-ray determination of the benzene crystal structure made by Cox and Smith,[37] more recent, and exciton wave functions, more general, than those used by Davydov. Their results are tabulated in Tables III and IV for exciton functions constructed from molecular states of symmetry B_{1u} and B_{2u}. The foregoing results belie our expectation of identifying the B_{1u} and B_{2u} band from polarization measurement because in each band the transitions are a, b, and c polarized in the same order of energy. The band widths are 720 and 58 cm^{-1}. The two optically active levels observed by Broude *et al.* with a separation

TABLE III. Symmetry and energy of states in the B_{1u} band and polarization and intensity of transition from the ground state to these states.

Representation in D_{2h}	Polarization of transition	Band splitting in cm^{-1}	Intensity ratio
B_{2u}	b	180	8×10^{-4}
B_{3u}	c	−180	17×10^{-4}
A_{1u}	forbidden	+360	0
B_{1u}	a	−360	29×10^{-4}

TABLE IV. Symmetry and energy of translationally invariant states in the B_{2u} band and polarization of transition from the ground state to these states.

Representation in D_{2h}	Polarization of transition	cm^{-1}
B_{3u}	c	−7
B_{2u}	b	+29
B_{1u}	a	−29
A_{1u}	forbidden	+7

of 25 cm^{-1} confirm the splitting qualitatively though not qualitatively and thus lend experimental support to the exciton nature of these bands in the benzene crystal.

(b) Naphthalene Crystal

Quantum mechanical treatment of the naphthalene crystal on the basis of the exciton theory has been carried out by Davydov and more recently by McClure and Schnepp.[38] The results are that the free molecular energy levels are split by crystal forces, each into α and β components such that transitions from the ground state to them are perpendicularly polarized. The exciton theory intensity figures for these transitions in terms of free molecular intensity are given in Table V.

The a- and b-axis polarized absorption spectra of

TABLE V.

Molecular	Symmetry F. group	Intensity a axis	Intensity b axis
B_{3u}	$A_u(\beta)$	0	0.045
B_{3u}	$B_u(\alpha)$	0.19	0
B_{2u}	$A_u(\alpha)$	0	0.76
B_{2u}	$B_u(\beta)$	0.1025	0

TABLE VI.

Upper electronic state	0—0 band cm^{-1} a pol	b pol	gas	F. group splitting cm^{-1}	Band shift cm^{-1}
B_{3u}	31 476	31 642	32 021	166	462
B_{2u}	33 783	33 610 or 33 460	35 910	170 or 320	2210 or 2290

single crystals of naphthalene were investigated by Prikhotjko,[39] Craig and Lyons,[40] and more recently by McClure and Schnepp. The latter workers made an attempt to interpret the observed structure of the a- and b-polarized spectra at 20, 77, and 298°K using single crystals of naphthalene of thickness 10 mm to 0.5 μ. Table VI depicts the observed effects of crystallization on the electronic transitions to the upper states in a naphthalene molecule. They are displacement and dichroic splitting of the 0—0 band in both transitions. The observed polarization behavior of the crystal absorption is described and compared with the exciton theory predictions in Table VII.

The fluorescence spectrum of the naphthalene crystal at 20°K, investigated in great detail by Obreimow and Sabaldas[41] and more recently by McClure and Schnepp, is polarized along the a and b axes of the crystal. Factor group splitting of vibrational level absent in the mixed crystal is evident in the doubling of the line 1384 cm^{-1} in the one and two vibrational quanta and of some

TABLE VII.

Observer	Region of study cm^{-1}	Temperature of study	Measured $I_b:I_a$ ratio	Exciton theory predictions $Ag-B_{2u}$	$Ag-B_{3u}$
Craig and Lyons	37 000 3150	Room temp.	4.5 5:1		
	37 000 3150	20°K	4.5 5:1	7.5:1	1:4.2
	0—0 band	20°K	100:1		
Craig	37 000 33 780		2—3:1		
McClure and Schnepp	37 000 33 780	20°K	Qualitative agreement with above		

[37] E. G. Cox and J. A. S. Smith, Nature **173**, 75 (1954).
[38] D. S. McClure and O. Schnepp, J. Chem. Phys. **23**, 1575 (1955).
[39] A. Prikhotjko, J. Phys. U.S.S.R., **8**, 257 (1944); Akad. Nauk U.S.S.R. Ser. Fiz **12**, 499 (1948).
[40] D. P. Craig and L. E. Lyons, J. Chem. Phys. **20**, 1499 (1952).
[41] W. Obreimov and K. Sabaldas, J. Phys. U.S.S.R. **7**, 168 (1943).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

TABLE VIII.

ν	Wave number in cm^{-1} Pure crystal	Mixed crystal	Vibrational wave number in cm^{-1} Pure crystal	Mixed crystal	Splitting
0	31 060.8	31 551.6	0	0	
1	29 681.2 29 673.0	30 174.2	1379.6 1387.8	1377.4	8.2
2	28 284.9 28 288.7	28 799.0	2765.9 2772.1	2752.6	6.2

combination lines involving 1384 cm^{-1} in the pure crystal. Measurements on the 1348 series are entered in Table VIII.

Thus the gross features of the absorption and the fluorescence spectra satisfy the predictions of the exciton theory for molecular crystals.

(c) Anthracene Crystal

The energy levels of the anthracene crystal were first treated by Davydov as an application of his theory of light absorption in molecular crystals. The absorption spectrum has been investigated by Obreimov and Prikhotjko[42] at 20°K, by Craig and Hobbins,[43] Lyons,[44] and Sidman.[45] The fluorescence spectrum has been studied by Obreimov, Prikhotjko, and Sabaldas[46] at 20°K, Pesteil and Barbaron[47] at 14°K, and Sidman[48] at 4°K. Craig and Hobbins, following Davydov, calculated the splitting, shift, and polarization ratio associated with an electronic transition in the anthracene crystal. Davydov's theory predicts that the second absorption system of anthracene will be split in the crystal by 16 000 cm^{-1} if it were due to a long axis transition and by 1000 cm^{-1}, if due to a short axis transition. Craig and Hobbins investigated the a- and b-polarized absorption spectra of the crystal and observed that the peak of the b spectrum in the second system occurred at 2680 A so that for $A_g - B_{3u}$ (long axis) assignment, the peak of the a spectrum should fall at 1900 A. Though they could not detect it, they observed a continuous rise of absorption until the spectrum was cut off at 2300 A in a crystal of thickness 0.1 μ so that the peak was expected beyond 2300 A. Recently, Lyons[44] reported that experiments with polarized light in the vacuum ultraviolet located the split components at 2200 A and at 1890 a.u. and the absorption is greater perpendicular to the b axis than parallel to it. This allows one to interpret the 1890 a.u. absorption as the a component of the vapor $A_g - B_{3u}$ transition polarized along the long molecular axis.

On the basis of the exciton theory, Craig and Hobbins calculated by first-order perturbation the polarization ratio of the a-b polarized spectrum in the second system and obtained theoretically molar extinction coefficients in the crystal from the solution values. The predicted value was 9100±500 for the b-crystal absorption, taking the solution extinction coefficient as 232 000 while the experimental value obtained by Bree and Lyons[47] was 9450. The polarization ratio ($I_b : I_a$) 1.8:1, observed by Craig and Hobbins for the first absorption system in anthracene, later confirmed by Ganguly *et al.*,[49] was quantitatively interpreted by Craig on the basis of exciton theory modified by crystal-induced mixing of two excited electronic states to be due to $A_g - B_{2u}$. This interpretation leads to a value 21 000 of molar extinction coefficient while the experimental value is 22 800 as measured by Bree and Lyons.

Thus, the observed behavior of electronic transition in the anthracene crystal confirms the predictions of the exciton theory. Calculations of Craig and Hobbins were made by considering the dipole-dipole interaction between molecules not more than 20 A apart. Craig showed that the influence of distant molecules made no difference in their calculated results because the interaction summation converges within a distance of 20 A in the anthracene crystal. Fox and Yatsiv[50] also critically discussed the dependence of the calculated splitting and shift on the shape, size, and direction of propagation of the exciton wave packet. They found that the observed splitting could not lead to an unambiguous long axis assignment for the second absorption system since by reasonable assumptions as to the shape and size of exciton wave packet, one may interpret the observed splitting on the basis of a short axis assignment while by these assumptions one may interpret the observed frequency shift as confirming long axis assignment.

The long wavelength absorption and fluorescence have been investigated in greater detail in the ab plane at 4°K by Sidman.[45] The absorption spectrum is found to consist of at least two regions, one which extends from 25 400 to 30 000 cm^{-1} in both the a- and b-polarized spectra and the other which extends downward from 2540 to 3000 cm^{-1} and appears only in the b polarization. The former, which consists of strong, broad absorption bands, the first band peak being located at 25 400±20 cm^{-1} by Sidman, at 25 380 cm^{-1} by Obreimov and Prikhotjko,[51] and at 25 350 cm^{-1} by Craig and Hobbins, as electronic vibrational transition to the two transitionally allowed levels of the free exciton band derived principally from L_a, B_{2u} state of an

[42] I. W. Obreimov and A. F. Prikhotjko, Physik. Z. Sowjetunion 9, 48 (1936).
[43] D. P. Craig and P. C. Hobbins; J. Chem. Soc. 2309 (1955).
[44] L. E. Lyons, J. Chem. Phys. 23, 1973 (1955).
[45] J. W. Sidman, Phys. Rev. 103, 96 (1956); J. Chem. Phys. 25, 115 (1956).
[46] Obreimov, Prikhotjko, and Sabaldas, J. Exptl. Theoret. Phys. (U.S.S.R.) 5, 1062 (1936).
[47] P. Pesteil and M. Barbaron, J. phys. radium 15, 92 (1954).
[48] J. W. Sidman, Phys. Rev. 102, 96 (1956).
[49] Ganguly, Chaudhury, and Mukherjee, 44th Session, Indian Science Congress (1957).
[50] D. Fox and S. Yatsiv, J. Chem. Phys. 24, 1103 (1956).
[51] I. W. Obreimov and A. F. Prikhotjko, Physik. Z. Sowjetunion 11, 948 (1936).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

anthracene molecule. The latter regions which consist of many weak bands and lines such as at 24 809, 24 836, and 24 989 cm^{-1} detected by Sidman is assigned by Sidman to transitions from the ground electronic state to the trapped exciton states of the crystal, predicted by Frenkel[12] and Davydov.[52] In addition to the phenomenon of propagation of a free exciton, Frenkel considered the case where the crystal lattice may suffer distortion in the neighborhood of a migrating exciton. His result is that the exciton may be trapped at the site of the local distortion produced by it, if the energy of the trapped exciton is less than that of the free exciton. The low intensity of absorption in this region is caused by a nonvertical Franck-Condon transition of the lattice, since the orientation of the molecule in the lattice is different in the upper and lower electronic states.

The fluorescence spectrum as observed by Sidman at 4°K consists of broad bands superimposed with sharp structures. The fluorescence spectrum has several origins, principally below 24 975 cm^{-1}, which do not coincide with the origin of strong absorption at 25 400 cm^{-1}. This gap of several hundred cm^{-1} between the origins of fluorescence and absorption has been verified by all workers and is significant. It shows that radiationless transference of excitation energy from free exciton levels to trapped exciton levels is rapid so that no fluorescence takes place from the free exciton levels at 25 400 cm^{-1}. This gap between the origins of absorption and fluorescence has been found to exist in crystalline naphthalene and phenanthrene and is similarly accounted on the basis of the trapped exciton theory of Frenkel. Sidman also points out that calculations of Craig[13] applied to Frenkel's prediction show that the relative shift of electronic transition in vapor mixed and pure crystals can be interpreted as a semiqualitative proof of the existence of trapped exciton.

(d) Naphthacene Crystal

The first absorption system of tetracene single crystals under light polarized parallel and perpendicular to the b axis of the crystal was investigated by Achis.[53] The maxima in one spectrum are displaced from those in other, so that Davydov splitting is compared. There seems to be no published absorption data on the second and third systems. Bree and Lyons[54] recently investigated the polarized ultraviolet spectrum of tetracene single crystals by the photoconductance method and confirmed splitting in the first absorption system as shown in Fig. 14. The flatness of the photoconductance curve in the second and the third absorption system does not give any information about splitting there. The absorption and photoconductance of the first system are entered in Table IX. Recently, the absorption and fluorescence spectra of tetracene in solid substitutional solutions at 20°K have been investigated by Sidman.[55] The lowest excited singlet state at 20 246 cm^{-1} has been found to be B_{2u}, L_a since the transition with the ground state is polarized along the short axis of the molecule.

TABLE IX.

Peak parallel to b direction	A[a]	19 235	20 715	22 320
	P[b]	19 160±50	20 750±50	22 270±300
Peak perpendicular to b direction	A	19 810	21 145	22 545
	P	19 860±50	21 190±50	22 370±400
Splitting between components	A	575	430	225
	P	700±100	440±100	

[a] A = absorption spectra.
[b] P = photoconductance curve.

(e) Phenanthrene Crystal

The a- and b-polarized absorption and fluorescence spectra of single crystal of phenanthrene have been investigated at 20°K by McClure.[56] The absorption spectra show that the strong a-polarized 28 613 cm^{-1} band and the strong b-polarized 28 660 cm^{-1} band act as if they are the two components of the 0—0 band split by intermolecular resonance. The former belongs to the B_u representation and the latter to the A_u representation of the factor group C_{2h} of the crystal so that the factor group splitting is 47 cm^{-1}. The vibrational frequency 1378 cm^{-1}, a C—C stretching mode also becomes split into 1370 and 1386 in the crystal. The total width of the exciton band is 280 cm^{-1}, the band beginning at 28 540 cm^{-1} and terminating at 28 820 cm^{-1}.

The weak absorption bands at 28 315 and 28 408 cm^{-1} on the red side of the 0—0 band are not regarded by McClure as a part of exciton band. Similar bands also occur in the crystals of anthracene and stilbene as observed by Sidman. The fluorescence originates at or near the weak absorption band at 28 315 cm^{-1}. Sidman has recently interpreted these low lying emitting levels as the trapped exciton bands of Frenkel.

(f) Hexamethyl Benzene Crystal

The molecular plane in the hexamethyl benzene crystal lies almost parallel to the ab plane of the crystal, the b axis of the crystal being in the plane of the benzene ring. The absorption spectra of the hexamethyl crystal under polarized light were investigated by Scheib *et al.*[57(a)] They concluded that the polarized light absorption in the near ultraviolet takes place predominantly

[52] A. S. Davydov, Izvest. Akad. Nauk S.S.S.R. Ser. Fiz. 12, 608 (1948).
[53] A. Y. Achis, quoted by A. S. Davydov, Akad. Nauk. Pamyati S. I. Vavilov, 197 (1952).
[54] A. Bree and L. E. Lyons, J. Chem. Phys. 22, 1630 (1954).
[55] J. W. Sidman, J. Chem. Phys. 25, 122–124 (1956).
[56] Donald S. McClure, J. Chem. Phys. 25, 481–485 (1956).
[57] (a) Scheibe, Hartwig, and Muller, Z. Electrochem. 49, 376 (1943). (b) O. Schnepp and D. S. McClure, J. Chem. Phys. 26, 83 (1957).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

in the plane of the benzene ring; though they report a definite absorption perpendicular to the ring planes, amounting to about ten percent of parallel absorption at room temperature. A detailed investigation of the absorption parallel and perpendicular to the ring plane was recently made by Schnepp and McClure[57(b)] at high resolution at 20°K in the 2800 A region. The hexamethyl benzene crystal contains one molecule per unit cell and because the excitation exchange energy between the neighboring molecules should be small, one can expect that the spectrum should be free from all the complications resulting from the presence of two or more molecules per unit cell and that the spectrum exhibits molecular properties and not crystal properties. The analysis shows that the 0—0 transition placed at 35 136 cm^{-1} is much stronger under light polarized perpendicular to the benzene ring than under light polarized in the plane. The in-plane absorption is due to a forbidden transition. Detailed quantum mechanical calculations show that the observed behavior, of the 0—0 band under polarized light is caused not by crystal effects but by the destruction of molecular symmetry by the unsymmetrical positions of the methyl groups in the S_6 structure. This distortion gives rise to weak perturbation of the π electron state by the perpendicular non-π electron states.

The fluorescence spectra, investigated by Schnepp and McClure at medium resolution, exhibit polarization of bands but not as complete as in absorption. As the 0—0 band, calculated from the fluorescence spectrum by use of Raman frequency, is located at 35 140±8 cm^{-1}, it is concluded that both absorption and fluorescence originate from 35 136 cm^{-1}. If this is so, then the crystalline band in this crystal must be either very narrow or the allowed level of the band must lie very near the bottom of the band.

Thus nonconducting excited states or exciton states do occur in most of the insulating aromatic crystals investigated as a result of excitation in the fundamental absorption band. The excited regions—called excitons—are believed to be mobile and capable of transporting energy through a crystal. The existing experimental evidence for exciton migration in these crystals rests mainly on indirect observations.

The work of Apker and Taft,[58,59] shows that excitons are able to create and destroy F centers. Exciton migration has been shown by Northrop and Simpson[30] to account quantitatively for the fluorescence quenching of aromatic hydrocarbons containing impurities; Sidman[45] has presented evidence derived from differences in the absorption and fluorescence spectra of anthracene which implies that excitons can be trapped at lattice imperfections. All these experiments depend on the interaction of excitons with particular centers of the lattice. That there occurs actual migration of excitons through the anthracene lattice has been recently demonstrated by the investigation of Simpson.[60] He devised a method which enabled him to measure the exciton motion directly. The observed diffusion length in the case of anthracene is found to be 460 A. In an isotropic medium this would correspond to a root-mean-square displacement of 1120 A between the points of origin and decay of an exciton.

Dexter[61] points out that a phosphor containing luminescence centers can be used as a detector of excitons. Thus it is possible to detect the diffusion of excitons from an activator free crystal into a contiguous crystal, if the latter contains centers capable of trapping the excitons. One surface of a thin specimen is covered by an exciton detector, the opposite face of the specimen is illuminated in the fundamental absorption band. The luminescence emitted by the detector is drived partly from excitation by that fraction of the primary radiation not absorbed in the specimen and partly from excitons originating in the specimen. A separate measurement of the transmission of the specimen suffices to separate the two contributions; hence to determine the flux of excitons into the detector.

The allowance for reabsorption of anthracene fluorescence by naphthacene, was found by measuring the emission from specimens which were made with the anthracene and detector deposited on separate glass plates, and then put together with a small air gap between. In this arrangement transfer by exciton diffusion was suppressed.

4. INFLUENCE OF CRYSTAL SIZE ON THE SPECTRAL NATURE AND DECAY TIME OF FLUORESCENCE

Though the matters discussed in the previous section bear testimony to the exciton nature of the electronic bands and to the migration of excitation energy by way of excitons, some experimental facts give evidence of energy migration by photon reabsorption. In this case efficiency of fluorescence escape from aromatic crystals shows a dependence upon crystal size, in contradiction with predictions of exciton theory.

The long-known fact that overlap of emission and absorption spectra of a system is apt to alter the spectral character of fluorescence due to self-absorption and re-emission, is given proper consideration in the case of organic crystals by the investigations of Kortum and Finckh,[62] Bowen and Lawley,[63] and Little and Birks.[64] They show that the transmitted photofluorescence spectrum of an organic crystal is critically influenced by crystal thickness. That this fact cannot be ignored is evident from the fluorescence spectrograms of anthra-

[58] L. Apker and E. Taft, Phys. Rev. **79**, 954 (1950).
[59] L. Apker and E. Taft, Phys. Rev. **82**, 814 (1951).
[60] O. Simpson, Proc. Roy. Soc. (London) **A238**, 402 (1956–1957).
[61] D. L. Dexter, J. Chem. Phys. **21**, 836 (1953).
[62] G. Kortum and B. Finckh, Z. physik. Chem. (Leipzig) **B52**, 263 (1942).
[63] E. J. Bowen and P. D. Lawley, Nature **164**, 572 (1949).
[64] W. A. Little and J. B. Birks, South African Association for the Advancement of Science Congress, Cape Town, South Africa (July, 1952).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

cene, *trans*-stilbene, *para*-terphenyl, and diphenylacetylene for thick crystals recorded by Koski,[65] Koski and Thomas,[66] and Sangster.[67]

Birks and Little[68] studied the spectral change in the fluorescence of each of these compounds as they changed the fluorescent specimen from single crystals of 1-cm thickness to microcrystalline layers deposited on a thin glass plate by evaporation of a dilute solution of the compound in xylene and the mode of observation from transmission to reflection. The fluorescence spectrum of anthracene (Fig. 9) gets extended into the ultraviolet as far as 3700 A and about 80% of this primary fluorescence suffers reabsorption in a one centimeter thick crystal. These observations were carefully re-examined by Birks and Wright[69], who used improved technique and made relative quantum intensity measurements on one centimeter thick crystal specimens and very thin microcrystalline layers deposited on quartz plates. They observed that the microcrystalline fluorescence spectrum corresponds to the complete molecular emission; 76% of the primary molecular emission from a 1-cm thick anthracene crystal could not escape reabsorption so that the probability of escape of molecular fluorescence, K was 24. In a thick *trans*-stilbene crystal 40% of the molecular fluorescdnce was reabsorbed.

These investigations lend support to the photon cascade theory formulated by Birks[70] and later on supported by Fowler and Roos.[71] Based on the fact that the absorption coefficient of an organic crystal is high, exceeding 10^3 in the first absorption band system of anthracene, and that there is overlap between the emission and absorption spectra, this theory proposes that in a thick crystal, the part of the primary molecular fluorescence which is emitted in the region of absorption is practically reabsorbed in a path length of a few wavelengths. This reabsorbed radiation is re-emitted as fluorescence, a fraction of which is again absorbed by succeeding layers, these processes of absorption and emission following each other until all the excitation energy either escapes from the crystal or is dissipated thermally by internal conversion. Thus the intermolecular energy transfer takes place by expansion of the excitation column, not by the mechanism of exciton migration but by the absorption and reabsorption of the primary fluorescence. This is an extension of Ganguly's[22] theory of transfer of energy from the anthracene lattice to the naphthacene molecules embedded in it through absorption of anthracene fluorescence by naphthacene molecules. But our recent measurement on the coefficient of absorption of naphthacene shows that the probability of naphthacene excitation by absorption of fluorescence of anthracene is remote.

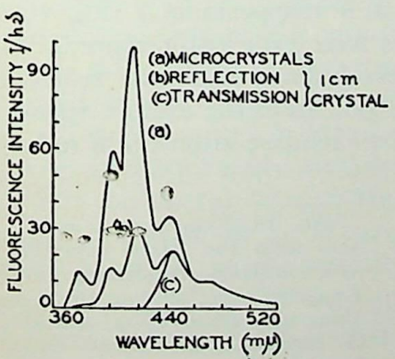


FIG. 9. Anthracene fluorescence spectra excited by 2537 A (after Birks and Wright).

The photon cascade theory seems to be strengthened by surprisingly few existing photofluorescence decay time measurements of organic crystals. Liebson[72] found that for anthracene the photofluorescence decay time t_f changed from 2.0±0.5 mμsec for a ½% solution in benzene to 17 mμsec, for a crystal. For terphenyl it is increased from 2.5±0.5 mμsec for a ½% solution to 11 mμsec for a crystal. For stilbene and diphenylacetylene crystals, they measured t_f to be 3.1 and 2.5 mμsec, respectively. The effect of crystal size on t_f was investigated by Birks and Little[73] who used the phase-shift method and observed t_f for thick single crystals exceeded by many times that for microcrystalline specimens. The data for anthracene and terphenyl represent two extreme cases. Whereas in the former the t_f increases from 3.5±1.0 mμsec for microcrystalline specimens to 14±2.0 mμsec for 1-cm³ single crystals, in the latter it changes from 3.5±1.0 mμsec for powder to 3.8±1.0 sec for 2-mm thick single crystals. Accounting for the observed increase in decay times in crystals by self-absorption and re-emission of the primary molecular fluorescence, they calculate the decay time t_f of technical fluorescence at 12.6±3.5 mμsec which agrees, within experimental error, with the experimental value 14±2.0 mμsec. They calculated the decay time $(t_f)_{00}$ of primary molecular fluorescence to be 3.1+0.8 mμsec in agreement with the experimental value 3.5±1 mμsec. Bose *et al.*[74] have observed that under x-ray excitation the decay times of anthracene and naphthacene emissions in a green anthracene crystal isolated by suitable filters were practically identical.

Splendid as the success of the photon cascade theory seems to be in interpreting the energy transfer mechanism in pure aromatic crystals, it is not so in mixed crystals. In fact, in mixed crystals it is more qualitative than quantitative and must be considered with caution. The basic fact utilized by the theory, *viz.*, that the primary photons emitted in the region of the first ab-

[65] W. S. Koski, Phys. Rev. **82**, 230 (1951).
[66] W. S. Koski and C. O. Thomas, J. Chem. Phys. **19**, 1286 (1951).
[67] R. C. Sangster, Massachusetts Institute of Technology Tech. Rept. No. 55 (1952).
[68] J. B. Birks and W. A. Little, Proc. Phys. Soc. (London) **A66**, 921 (1953).
[69] J. B. Birks and G. T. Wright, Proc. Phys. Soc. (London) **B67**, 657 (1954).
[70] J. Birks, *Scintillation Counters* (Pergamon Press, London; McGraw-Hill Book Company, Inc., New York, 1953).
[71] J. M. Fowler and C. E. Roos, Phys. Rev. **98**, 996 (1955).
[72] S. H. Liebson, Nucleonics **10**, 41 (1952).
[73] J. B. Birks and W. A. Little, Proc. Phys. Soc. (London) **A66**, 921 (1953).
[74] H. N. Bose *et al.*, Proc. Natl. Inst. Sci. India **18**, 397 (1952).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

sorption band system of a crystal are reabsorbed completely in this region due to very high absorption coefficients (10^2-10^3) is violated in the case of energy transfer from the host lattice to impurity molecules in mixed crystals. In spite of very high absorption coefficients of impurity molecules the effective absorption by a mixed crystal in the region of impurity absorption is much short of being total, unlike in pure crystals, because of the very low concentrations ($10^{-4}-10^{-5}$) mole solute/mole solvent at which the transfer of energy is almost total. The fluorescent spectrum studied by reflection, instead of transmission, to eliminate the small absorption effect of a thin microcrystalline layer of anthracene containing low percentage of naphthacene deposited from xylene solution on a fused quartz plate exhibits the naphthacene fluorescence bands in high intensity.[75] The result of drawing a parallelism between absorption effects due to the host lattice and impurity molecules is evident from the following considerations.

The relation

$$q_{ex}=\frac{1}{1+(K_y[Y]/K_{ex})}=\frac{1}{1+K[Y]},$$

where q_{ex} = technical fluorescence quantum efficiency of the host crystal X, K_y = probability of photon capture by the impurity y, K_{ex} = probability of photon escape in X crystal fluorescence, and $[Y]$ = concentration of Y, derived on the photon cascade theory in the absence of internal quenching, is similar to $N_x=1/(1+K[Y])$, where N_x = technical quantum efficiency of X-crystal luorescence, K = exciton capture probability of Y elative to that of X, derivable from the exciton theory. In both of them the important quantity is the energy transfer parameter k which determines the efficiency of the process. The value of k determined by Bowen *et al.* from the experimental N_x *vs* $[Y]$ curve for anthracene-contaminated naphthalene crystals excited by 2900 to 3000 A was 5×10^4 and the value obtained by Wright who used 2540 A as the exciting radiation was 4×10^4. In the relation $K_{ex}+K_x+K_y$, $[Y]=1$ where K_x = photon capture probability of X, given by the cascade theory, putting $Y=2.5\times10^{-5}$ mole, anthracene per molar naphthalene for equal intensities of the components and assuming the most favorable but experimentally unrealized condition $K_y=1$, we have $K_{ex}+K_x+2.5\times10^{-5}=1$. As the theory assumes no interaction between the host and the guest molecules, K should not differ from its value in a pure crystal. Substituting $K_x=0.9565$ from scintillation data which always gives a somewhat larger value, we have $k_{ex}=4.3475\times10^{-2}$ and $k=23$ far away from the experimental value. The only support for the theory in mixed crystals is Bowen and Mikiewicz's[76] result that the intensities of anthracene naphthacene fluorescence components become equal at 1.8×10^{-5} g naphthacene/g anthracene in anthracene-naphthacene crystals and at 9.10^{-5} g/g in solution, giving a ratio 0.2. The explanation that this difference results from the reduction of the effective absorption coefficient in solutions due to random alignment by a factor 0.25 relative to that in crystals where the Y molecules align parallel to X molecules applies to all such crystals and is not supported by the experimental facts that light in a double refracting crystal is bifurcated into two perpendicularly polarized components and that the absorption coefficients of a substance such as anthracene in solution are greater[77] than those in crystal by a factor ~8. Moreover, the disagreement by 25% is glaring.

5. SCINTILLATIONS FROM ORGANIC CRYSTALS

The scintillation property of fluorescent organic crystals has made them of great use in scintillation counting of nuclear radiations, but there seems to have been little attempt to understand its nature until the recent work of Taylor *et al.*,[8] Franzen *et al.*, Hopkins, Birks, etc.[78–82] We review the fundamental facts in the scintillation process because any satisfactory theory of the fluorescent process in these crystals must not ignore them. That the energy transfer mechanism in organic crystals under photon or corpuscular excitation is qualitatively independent of the mode of excitation seems to be borne out in Birk's[83] experiments. Using crystals of naphthacene containing different concentrations of anthracene he investigated their scintillation response to alpha particles. The intensities of naphthalene (N) and anthracene (A) emissions were measured separately by introducing suitable filters between the crystals and the photo tube. His plots of (N) and (A) fluorescence intensities, relative to pure crystal, against A concentrations, show that not only the N fluorescence is progressively suppressed with A concentration but also the quantity suppressed in N reappears in A (Fig. 10). The experimental curves were excellently reproduced in the theoretical curves obtained from the relation $n=1/(1+kx)$ derived by him from the exciton theory of energy transfer where n = relative intensity of radia-

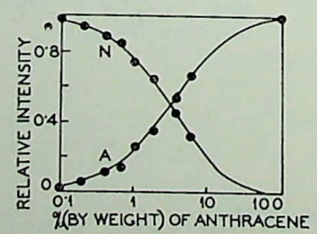


FIG. 10. Fluorescence emission from the mixed crystals of naphthalene containing anthracene as a function of anthracene concentration. A and N represent anthracene and naphthacene fluorescence, respectively (after Birks).

[75] N. S. Bayliss, Rev. Pure Appl. Chem. 1, 64 (1951).
[76] E. J. Bowen and E. Mikiewicz, Nature 159, 706 (1947).
[77] N. S. Bayliss, Rev. Pure Appl. Chem. 1, 64 (1951).
[78] Franzen, Peele, and Scherr, Phys. Rev. 79, 742 (1950).
[79] Taylor, Remley, Jentschke, and Kruger, Phys. Rev. 83, 169 (1951).
[80] Frey, Grim, Preston, and Gray, Phys. Rev. 82, 372 (1951).
[81] Jentschke, Eby, Taylor, Remley, and Kruger, Phys. Rev. 82, 170 (1951).
[82] J. I. Hopkins, Rev. Sci. Instr. 22, 29 (1951).
[83] J. B. Birks, Proc. Phys. Soc. (London) A63, 1044 (1950).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

tions, k = the exciton capture probability of an anthracene molecule relative to that of a naphthalene molecule, x = anthracene concentration in moles per mole of naphthalene; with $k=43$ for α particles. These curves are similar to those obtained by Bowen *et al.*[84] and Wright *et al.*[85] in case of photofluorescence of anthracene in naphthalene with $k=5\times10^4$.

A similar relation was derived by Birks and Black[86] on the exciton to account successfully for the irreparable deterioration in anthracene scintillation response produced by a prolonged intense α-particle bombardment so as to cause a permanent molecular damage as reported by Birks[87] and Wright.[88] They regard the damaged molecules as quenching centers just as quenching impurity molecules in a host lattice and found a value 1000 for k which is the exciton capture probability of a damaged molecule relative to an undamaged anthracene molecule.

The investigations of the scintillation response S of anthracene to protons of $E\leq17$ Mev by Franzen *et al.*,[78] Taylor *et al.*,[79] and Frey *et al.*[80] to electrons of E from 1 Kev to 3 Mev by Jentschke *et al.*[81] and Hopkins,[82] to α particle of $E\leq2.1$ Mev by Birks[87] and Taylor *et al.*,[79] to deuterons of $E\leq11$ Mev by Taylor *et al.*, show that the response S of anthracene to an ionizing particle of energy E and residual range r, depends on the nature of the particle and is proportional to E above a certain energy level. Measurements of the relative response of other organic crystals such as naphthalene,[87] stilbene,[79] terphenyl,[89] etc., and of organic solutions[90] to different particles show that they behave in the same manner as anthracene so that the behavior of anthracene can be regarded as typical of the organic substances studied. The $S-E$ curves of anthracene for different ionizing radiations are shown in Fig. 11(a). Birks' investigations[91] deserve special mention in that, unlike others, he compared the scintillation response S of anthracene and other crystals to different ionizing particles, by considering not the $S-E$ curve but the variation of the specific fluorescence ds/dr (expressed in arbitrary units of volts/cm air equivalent) with the specific energy loss dE/dr (measured in Mev/cm air equivalent) where r (measured in air equivalent) is the residual range of the particle within the crystal. He showed that the nonlinear portion of the $S-E$ curves was due to the nonlinear variation of the specific fluorescence with the specific energy loss. At low values of dE/dr (electrons of $E>125$ Kev) ds/dr increases linearly with correspondence to S being proportional to E. At high values of dE/dr (α particles of $E<5$ Mev) ds/dr is practically constant so that the scintillation S is proportional to the residual range r of the particle rather than to its energy E. This linearity in the $S-r$ curves for anthracene, terphenyl, and stilbene was found to hold good for r about 8 mm. The ds/dr *vs* dE/dr curve for anthracene and $S-r$ curve for stilbene are shown [Fig. 11(b) and Fig. 12].

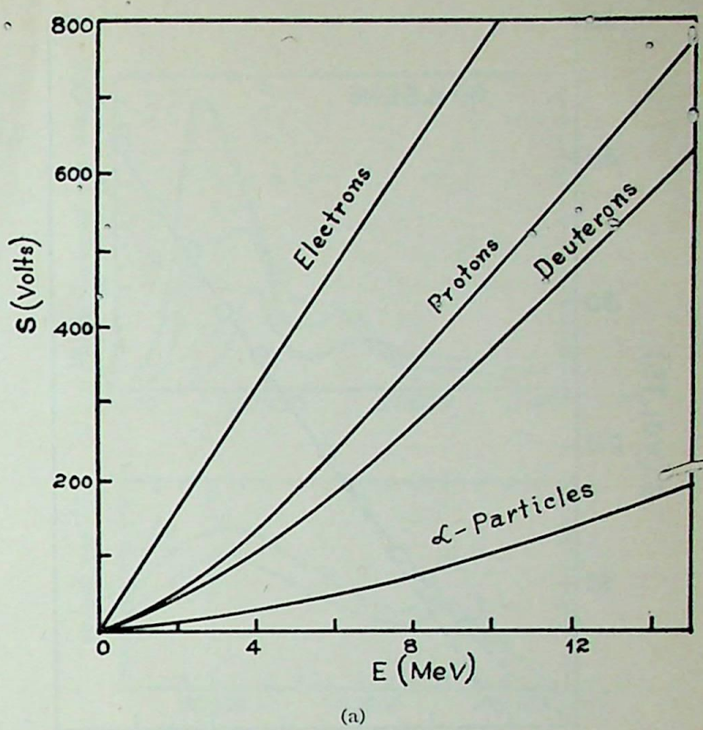


FIG. 11. (a) Relative scintillation intensity (S) of anthracene, produced by different ionizing particles, as a function of their energy (E) (after Jentschke *et al.*). (b) Specific fluorescence *vs* specific energy loss for anthracene (after Birks).

The exciton theory was extended by Birks[92] to account for the foregoing experimental facts. The passage of an ionizing particle through the crystal produces along its path a local concentration of damaged or ionized molecules and he regards them as quenching agents for the excitons produced by the ionizing particle. The linear density of excitons produced and the local concentration of nonradiative molecules at any point on the particle track are both proportional to the specific energy loss of the particle and can be put equal to $A(dE/dr)$ and $B(dE/dr)$ molecules per undamaged molecule, respectively. Denoting by k the exciton capture probability of a damaged molecule relative to

[84] E. J. Bowen, Nature **142**, 1081 (1938); E. J. Bowen and E. Mikiewicz, Nature **159**, 706 (1947); E. J. Bowen and P. D. Lawley, Nature **164**, 572 (1949); Bowen, Mikiewicz, and Smith, Proc. Phys. Soc. (London) **A62**, 26 (1949).
[85] G. T. Wright, Proc. Phys. Soc. (London) **A66**, 777 (1953).
[86] J. B. Birks and F. A. Black, Proc. Phys. Soc. (London) **A64**, 511 (1951).
[87] J. B. Birks, Proc. Phys. Soc. (London) **A63**, 1294 (1950).
[88] G. T. Wright, Proc. Phys. Soc. (London) **A66**, 777 (1953).
[89] R. Hofstadter, Nucleonics **6**, 72 (1950).
[90] Reynolds, Harrison, and Hill, Phys. Rev. **82**, 317 (1951).
[91] J. B. Birks, Proc. Phys. Soc. (London) **A64**, 874 (1951).
[92] J. B. Birks, Phys. Rev. **84**, 354 (1951).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

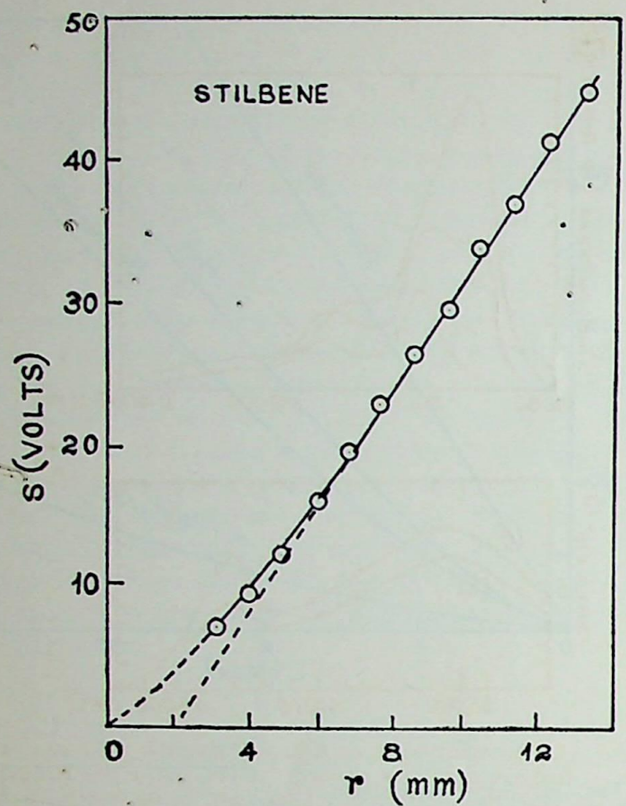


FIG. 12. Scintillation response of stilbene to α particles as a function of their air range r (after King and Birks).

an undamaged one, he deduces the relation,

$$\frac{ds}{dr}=\frac{A(dE/dr)}{1+KB(dE/dr)}, \tag{49}$$

which is reduced to $ds/dr=A(dE/dr)$ at low values of dE/dr and to $ds/dr=A/KB=$constant, at high values of dE/dr in agreement with observed facts. The experimental values of dE/dr and ds/dr used for substitution in Eq. (49) were obtained from the observations of Birks,[87] Curie,[93] Livingstone and Bethe,[94] and Bethe[95] and gave for $A=82.5$, $KB=7.15$ in anthracene.

The variation of dS/dr with dE/dr calculated from (49) agrees well with the experimental curve obtained for anthracene by Jentscke *et al.*[81] as shown in Fig. 11. The $S-E$ curve for electrons, protons, α particles, etc., calculated from (49) are in excellent agreement with those obtained experimentally.[96]

To account for the observed nonlinearity in the $S-r$ curves for $r<8$ mm, Birks[96] introduced a multiplying factor ϕ in (49) in consideration of the different behavior of excitons produced near the crystal surface. The function ϕ is given by

$$\phi=1-\tfrac{1}{2}\left\{\exp\left(-\frac{r}{a_0}\right)-\left(\frac{r}{a_0}\right)E_i\left(\frac{r}{a_0}\right)\right\},$$

where a_0—mean free path of an exciton and $E_i(r/a_0)$ exponential integral. The experimental $\phi-r$ curve for electrons in anthracene agrees well with the theoretical one.

The absolute efficiency of scintillation, i.e., the total energy of fluorescence emission/total energy of scintillation particles, of anthracene was measured by Birks and Szendrei[97] to be $3.76(\pm0.07)\%$ for thick anthracene crystals at room temperature, excited by fast electrons. This corresponds to an energy expenditure $E=70.5(\pm3.8)$ ev per fluorescence photon. The cascade theory formulated by Birks gave $E=68$ ev per photon in agreement with the experimental value.

Liebson[72] investigated the temperature dependence of the scintillation response S and the scintillation decay time t for anthracene from 103° to 283°K. His results show that a linear relationship between $\log_{10}S$ and t_1 holds good. Birks[98] worked out the relationship between S and t_1 on the photon cascade theory in $\log S=\log A+Bt_1$ where A is a constant and $B=\log q_0 t_0$, q_0 being the quantum efficiency of photon emission, occurring in cascade and of molecular decay time t_0. Substituting the experimental value of B from the graph and $q_0=0.9$, Birks found $t=3.8$ mμsec for the molecular decay time of anthracene at room temperature in excellent agreement with the experimental value 3.5 ± 1.0 mμsec obtained by Birks and Little.[73]

Sangster and Irvine[99] have studied the scintillation and photofluorescence behavior of a large number of crystalline organic compounds and the utility of Birks's photon cascade theory of the scintillation process has been demonstrated.

6. PHOTOCONDUCTIVITY OF ORGANIC CRYSTALS

Byke and Bork,[100] and Volmer[101] observed that anthracene crystals begin to photoconduct under the irradiation with wavelengths shorter than 4000 A. Later works of Putseiko, Akamato and Inokuchi, Vartanyan, Inokuchi, Eley *et al.*,[102] Mette and Pick,[103] Northrop and Simpson,[104] and Lyons and Morris[105] have confirmed that not only photoconduction but also semiconduction takes place in some aromatic crystals. This photoconductivity effect raises the question whether production and migration of holes have anything to do with the transport of excitation energy

[93] Mme. P. Curie, *Radioactivity* (Herman et Cie, Paris, 1935).
[94] M. S. Livingston and H. A. Bethe, Revs. Modern Phys. 9, 245 (1937).
[95] H. A. Bethe, Revs. Modern Phys. 22, 213 (1950).
[96] J. B. Birks and J. W. King, Phys. Rev. 86, 569 (1952).
[97] J. B. Birks and M. E. Szendrei, Phys. Rev. 91, 197 (1953).
[98] J. B. Birks, Phys. Rev. 95, 277 (1954).
[99] R. C. Sangster and J. W. Irvine, Jr., J. Chem. Phys. 24, 670 (1956).
[100] Byke and H. Bork, Ber. deut. physik. Ges. 8, 621 (1910).
[101] M. Volmer, Ann. Physik 40, 775 (1913).
[102] E. K. Putseiko, Doklady Akad. Nauk S.S.S.R. 71, 641 (1949); A. T. Vartanyan, *ibid.* (1950); H. Akamato and H. Inokuchi, J. Chem. Phys. 18, 810 (1950); H. Inokuchi, Bull. Chem. Soc. Japan 29, 131 (1956); Eley, Parfitt, Perry, and Taysum, Trans. Faraday Soc. 49, 79 (1953).
[103] H. Mette and H. Pick, Z. Physik 134, 566 (1953).
[104] D. C. Northrop and O. Simpson, Proc. Roy. Soc. (London) A234, 124 (1956).
[105] L. E. Lyons and G. C. Morris, Proc. Phys. Soc. (London) B69, 1162 (1956).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

through the anthracene lattice, even though want of any evidence of ionization in the anthracene lattice led Franck and Livingston[9] to discard the electron migration mechanism. Recent studies of the photoconductivity of anthracene have been made by Carswell,[106] Carswell and Lyons,[107] Chynoweth and Schneider,[108,109] and Bree, Carswell, and Lyons.[110] Carswell and Lyons' experimental arrangement uses a single anthracene crystal sublimed in a CO_2 atmosphere, mounted with Aquadag on platinum electrodes and irradiated with monochromatic radiation. Their important observations are: (a) plot of photocurrent *vs* frequency[106] of exciting radiations closely reproduces the optical absorption spectrum (Fig. 13), (b) the ratio of the current produced with light polarized parallel to the *b*-crystal axis to that parallel to the *a*-crystal axis was 1.5, using a mercury 3650 A light source[111]; (c) the photocurrent is proportional to the intensity of light; (d) quantum efficiency is approximately 10^{-4}, under his experimental conditions, constant over the near ultraviolet, and nearly equal to the fluorescence quantum efficiency; (e) the intentional incorporation of naphthacene as an impurity in the anthracene lattice so as to change the relative heights of the 5325 A and 4425 A fluorescent bands to 37:1 apparently affects neither the photocurrent nor the time constant.[111] The investigations of Chynoweth and Schneider[109] are of great significance in that their results support the adoption in organic crystals of the conventional energy band picture usually associated with inorganic crystals. The fact that irradiation of anthracene crystals within the fundamental absorption spectrum cannot liberate electrons completely,[112] combined with their observation that the majority of the carriers are positive holes suggests to them the following mechanism: light absorption raises the electrons either directly or through the intermediate process of exciton formation, into a metastable level from which they can either recombine with their holes, emitting fluorescence, or are trapped in a relatively stable level. Under the action of the field the holes drift through the lattice, but the drift velocity depends upon hole traps of different depths. Under steady-state conditions the rate of creation of carriers equals the rate of recombination between positive holes in the full band and the trapped electrons, as also the rate of trapping of carriers equals the rate of detrapping.

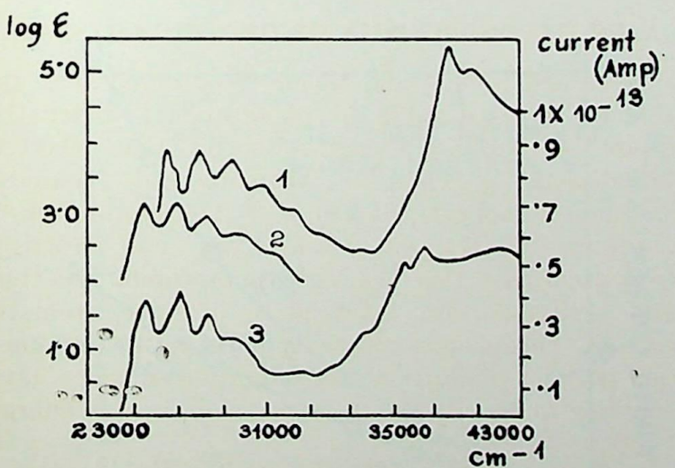


FIG. 13. (1) Solution absorption spectrum of anthracene. (2) Crystal absorption spectrum of anthracene. (3) Spectral photocurrent curve in anthracene crystal. Plot of photocurrent *vs* frequency of exciting radiations showing reproduction of the optical absorption spectrum (after Carswell).

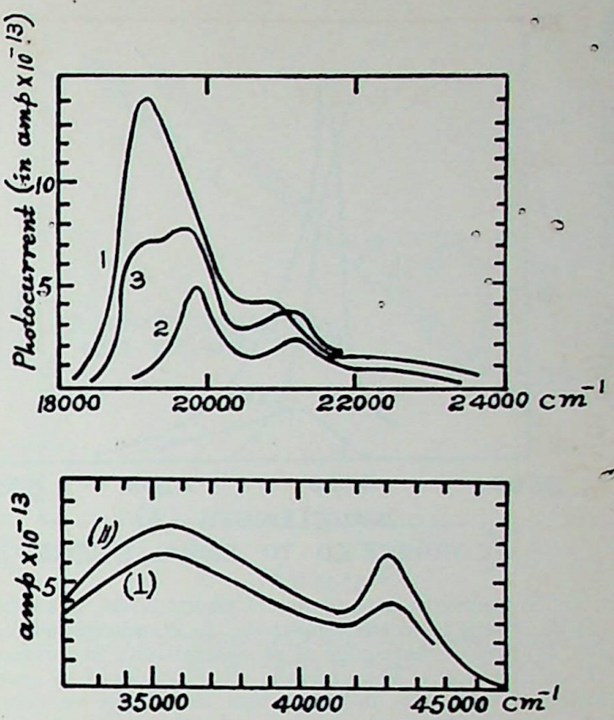


FIG. 14. Photocurrent in naphthacene single crystal as a function of frequency, in two ranges of wavelengths. In 1, light is polarized $||b$; in 2, $\perp b$; in 3, unpolarized. In the lower diagram "||, ⊥" are with respect to a direction which makes 12° with the *b* crystallographic axis (after Bree and Lyons).

If this mechanism is to be adopted for the emission of characteristic fluorescence of anthracene, it remains to be seen how it explains the want of reproduction of some photoconductivity phenomena in corresponding fluorescent phenomena as noted in (d) above. Leaving aside the question of transport of energy through the lattice of anthracene, the observation (e) made by Carswell suggests that immediate transfer of energy to naphthacene or pentacene in an impure anthracene crystal takes place by a mechanism other than hole migration.

The correspondence between the optical absorption spectrum and the photoconductance curve of an anthracene crystal, as observed above, led Bree and Lyons[54] to study the ultraviolet polarized spectrum of tetracene

[106] D. J. Carswell, J. Chem. Phys. **21**, 1890 (1953).
[107] D. J. Carswell and L. E. Lyons, J. Chem. Soc. 1734 (1955).
[108] A. G. Chynoweth and W. G. Schneider, J. Chem. Phys. **22**, 1021, (1954).
[109] A. G. Chynoweth, J. Chem. Phys. **22**, 1029 (1954).
[110] Bree, Carswell, and Lyons, J. Chem. Soc. 1928 (1955).
[111] D. J. Carswell (private communication to S. C. Ganguly, 1954).
[112] E. J. Bowen, Nucleonics **10**, 14 (1952).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

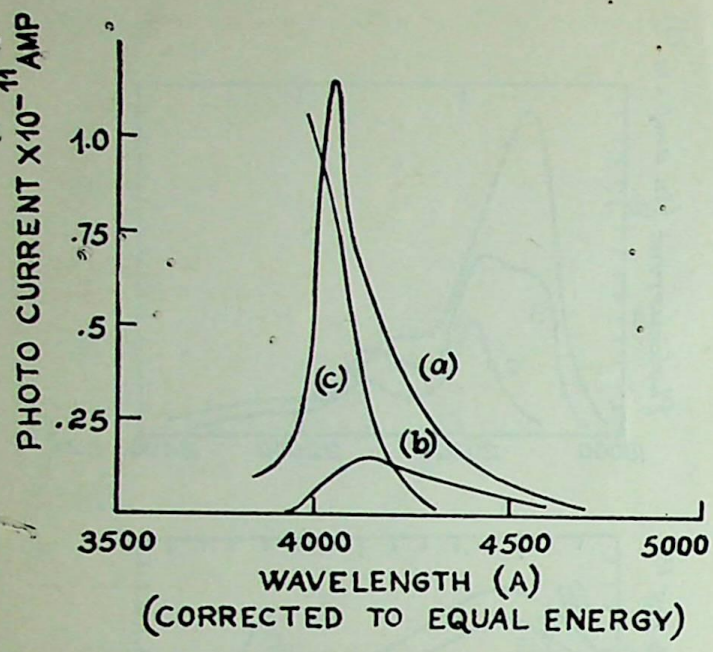


FIG. 15. Wavelength dependence of photocurrent in sandwich cell; (*a*) illuminated electrode positive, (*b*) illuminated electrode negative, (*c*) approximate edge of photosensitivity in surface cell (after Compton *et al.*). The observed effect produced by reversal of polarity of electrodes indicates that the charge carriers are mostly positive holes.

single crystals by the photoconductance method. They observed that the first system is definitely *b*-axis polarized and exhibits Davydov splitting (Fig. 14) as discussed in Sec. 3. Lyons and Morris[105] investigated the spectral dependence of photocurrent in naphthalene, phenanthrene, diphenyl, and *p*-terphenyl and compared it with the solution absorption spectra in the compounds. The photoconductivity takes place in the absorption region and not outside it. The measurements of Riehl,[113] Bree, Carswell, and Lyons[110] suggest that there is no true bulk photocurrent in anthracene and that the photoconductivity in anthracene is a surface phenomenon. This led Lyons[114] to explain the spectral dependence of photoconductivity in a surface cell on the assumption that the light energy absorbed by the crystal creates excitons, and that only those excitons that can migrate from their points of origin to the surface layer can generate charge carriers. The number n of excitons reaching the surface is given by

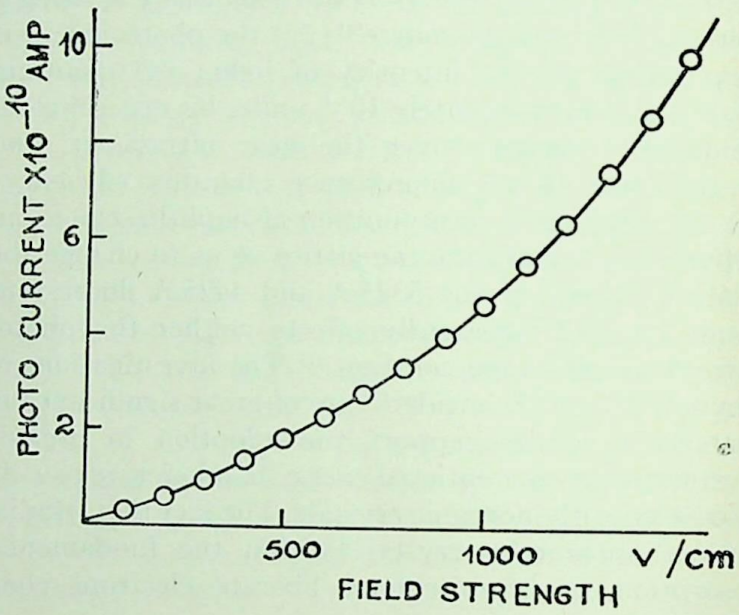
$$n=KI_0\left[\frac{E}{1-q}-\frac{E_q^2}{2}\left\{2(1-q)^2+\frac{1}{1-q}\right\}+E^3\right], \quad (50)$$

where E denotes the optical density of a molecular layer, I is the intensity of the incident light, q is the probability that an exciton crosses a layer of molecules, and k is a constant depending upon the specimen. If E is small, then neglecting higher powers of E, $n=KI_0E/(1-q)$ and the photocurrent is given by

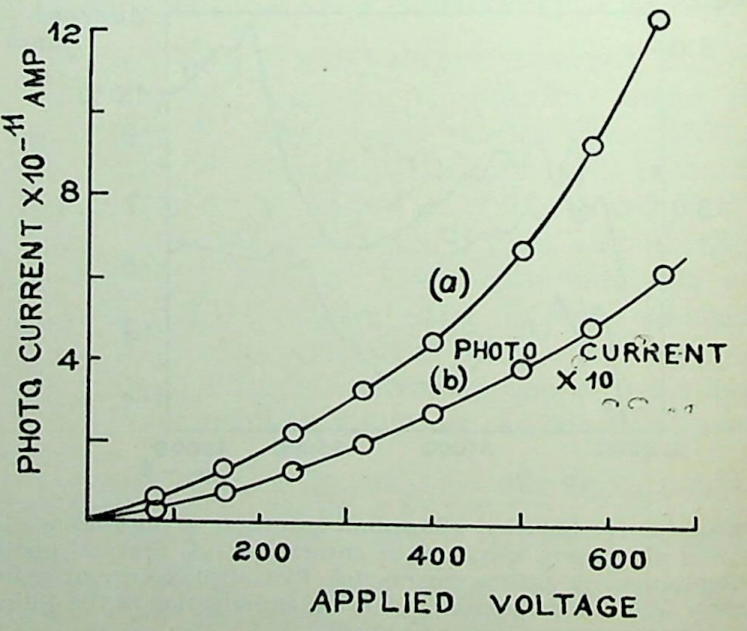
$$i=e\mu V=K'I_0VE/(1-q), \quad (51)$$

where μ is the mobility of the charge carrier and V is the applied voltage. The photocurrent is linearly dependent on (a) the optical density of the crystal, (b) the incident light intensity, and (c) the applied voltage. The linear dependence on the incident light intensity and on the applied voltage was demonstrated by the investigations of Carswell and Lyons[107] and Chynoweth and Schneider,[108] respectively. Lyons[114] showed that, in the case of the tetracene crystal, the plot of optical density of the crystal against photocurrent was a straight line and thus exhibited the predicted dependence.

A recent detailed investigation of surface and bulk photoconductivity of anthracene was undertaken by Compton, Schneider, and Waddington.[115] They used three types of crystals, namely, large scintillation crystals, thin layer crystals grown from dimethyl-



(a)



(b)

FIG. 16. (a) Photocurrent as a function of field strength (surface cell); (b) Photocurrent as a function of applied voltage (sandwich cell), (*a*) illuminated electrode positive, (*b*) illuminated electrode negative, crystal thickness 2 m (ordinate scale values are multiplied by ten) (after Compton *et al.*).

[113] N. V. Riehl, Zhur. Fiz. Khim. **29**, 959 (1955).
[114] L. E. Lyons, J. Chem. Phys. **23**, 220 (1955).
[115] Compton, Schneider, and Waddington, J. Chem. Phys. **27**, 160 (1957).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

formamide solutions, and sublimation flakes, and practically got rid of the space-charge effect, observed by earlier workers, by using electrodes of colloidal graphite, evaporated metals, and silver paste to secure better contacts with the crystal. To investigate the bulk photocurrent, they constructed a sandwich cell fitted with a guard ring on the surface separating the two electrodes. [One of the electrodes, called the front electrode, which was transparent, conducting and receiving the exciting light, was connected to a stabilized source of power; the other electrode, called the rear electrode, was connected to a microammeter.] If the guard ring is earthed, surface current is prevented from reaching the rear electrode.

Besides confirming the reported spectral dependence of surface photoconductivity and that the photocurrent is mostly caused by the mobility of positive holes, they arrived at the following results:

(1) Anthracene crystals exhibit a true bulk photoconductivity which is not affected by the nature of ambient gases. The bulk photocurrent in an anthracene sandwich cell is much less than the photocurrent in a surface cell under comparable conditions. The spectral dependence of bulk photocurrent is unlike that of surface photocurrent in that the maximum photocurrent occurs on the long wavelength side of the anthracene absorption band (Fig. 15).

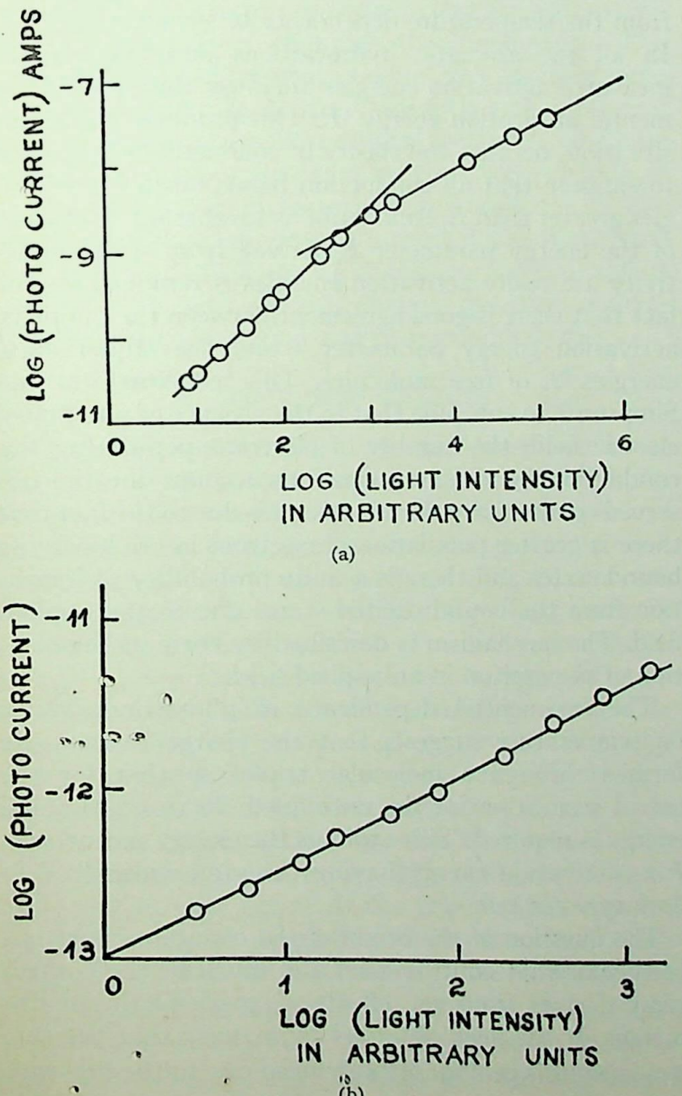


FIG. 17. Photocurrent as a function of light intensity (a) surface cell and (b) sandwich cell (after Compton *et al.*).

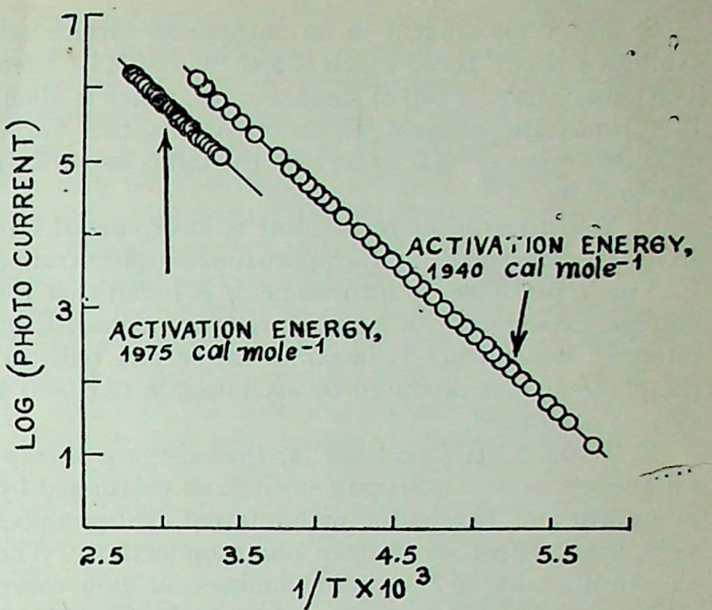


FIG. 18. Photocurrent as a function of temperature (after Compton *et al.*).

(2) The surface and bulk photocurrent in anthracene, generally nonohmic, exhibits an ohmic dependence on applied voltage at very low strength (less than 60 v cm^{-1} in a surface cell) and increases at a rate higher than linear at higher field strengths. The photocurrent, i, showed no sign of saturation, even with 25 kv cm^{-1} in a surface and 20 kv cm^{-1} in a sandwich cell (Fig. 16). Plots of V/i against V, in the case of a surface cell were straight lines, indicating that i is proportional not to V, as demanded by Lyons' theory, but to $V/(a-bV)$.

(3) In the surface cell the photocurrent shows a linear dependence on the light intensity except at very high level of intensity, whereas in a sandwich cell it varies as the square root of the light intensity (Fig. 17).

(4) The photocurrent displays an interesting dependence on crystal orientation; it is greatest in the b direction, about ten times smaller in the a direction, and one-hundred times so in the c direction. This indicates that the charge carriers are more mobile in the b direction than in the a and c directions by factors 10 and 100, respectively.

(5) The photocurrent studied in a surface cell or in a sandwich cell over a temperature range from 55 or 70°C to −130°C in an atmosphere of pure dry argon, follows the relation

$$i = i_0 e^{(-E/KT)},$$

where $E = 0.172 \pm 003$ ev in a surface cell and 0.176 ± 010 ev for a sandwich cell (Fig. 18).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

(6) The photocurrent in an anthracene surface cell exhibits a decay time, which is less than 30×10^{-5} sec. The decay shows a tail of some seconds, which is about 10^{-3} times the original photocurrent. As this varies markedly from crystal to crystal, the tail is most likely due to flaws.

(7) When tetracene or acridine is incorporated into an anthracene crystal in the approximate concentration 10^{-3} mole per mole of anthracene, it is found that the voltage dependence of photocurrent is unaltered, its intensity dependence is linear in the surface cell, and the photocurrent produced by excitation in the 3800 A system is reduced.

(8) There is little evidence for the existence of trapping levels in an anthracene crystal, as postulated by many workers. Irradiation with infrared light produces very small effect on surface photoconductivity. The electrical analog of the thermoluminescent glow curve shows that the concentration of traps of depth from 0.15 to 0.5 ev is less than 10^8 cm^{-3} in a pure anthracene surface or sandwich cell, and this is not dependent on the nature of the ambient gas. These experimental results are explained on the following model: The carriers of charge, which are positive holes, are generated in the bulk as well as on the surface, the carrier mobility being greater in the latter case than in the former. Lyons' calculations can be applied with the modification that q denotes the probability of charge carrier crossing a molecular layer. The photocurrent thus remains proportional to the optical density as observed by Lyons but the smaller the optical density is, the greater is the proportion of carriers formed at a good depth below the surface. This gives rise to the long wavelength tail of surface current in thick crystals and becomes of great importance in a sandwich cell. The photocurrent in the bulk is recombination limited, i.e., the charge carriers (positive holes) end their life by recombination before reaching the electrodes.

In the surface cell these carriers migrate to the surface along the shortest route, but because the surface in a real crystal is not perfectly flat, this route at many points will not be perpendicular to the field direction. Thus, the probability q of a charge carrier crossing a molecular layer is affected by V and if the phenomenon is averaged over the distance between electrodes, q can be expressed as $q=a+bV$. The equation for i becomes

$$i=\frac{KVI_0E}{1-a-bV}=\frac{KVI_0E}{a'-bV},$$

which explains the experimental fact that V/i depends linearly on V. The mechanism of formation of charge carriers in anthracene and other organic crystals is not clearly understood. Though photoconductivity is produced by absorption of light, it is not a primary product of light absorption because the first excited states of anthracene crystal are exciton states where the electrons and holes, created by the act of light absorption, migrate remaining bound to each other and produce no photocurrent. The experimental facts that the quantum efficiency for photoconductivity is of the order of 10^{-4}, and that the photocurrent is profoundly influenced by impurities (as observed by Carswell and Lyons[107] and Compton *et al.*[115]), and reduced by neutron irradiation (as observed by the latter group), cannot be explained on such an assumption.

In many photoconductors in which infrared irradiation has little effect on the photocurrent, optical and thermal formation of charge carriers requires identical energies and the absorption edge occurs corresponding to the energy E in the equation for semiconductivity $\sigma=\sigma_0e^{-E/2kT}$. This relation is obeyed by the aromatic[113] hydrocarbons such as anthracene, naphthacene, pentacene, etc. If the generation of charge carriers is not a primary product of light absorption in anthracene, one cannot expect that the energy parameter E would have the energy value corresponding to the absorption edge. In fact, the energy 3.2 ev associated with an exciton in the 3800 system is greater than the activation energy of semiconductance 1.63 ev as found by many workers[103,104] from the temperature dependence of semiconductivity. In all the aromatic hydrocarbons investigated, the measured activation energies are lower than the fundamental absorption energy ^1E. This produces a difficult situation, because the theory of conductivity leads one to suppose that all conduction bands must have energies greater than E. Any doubt as to whether the values of the energy parameter E derived from semiconductivity are really activation energies is removed by the fact that there is good agreement between the impurity activation energy parameter ϵ and the triplet state energies ^3E of free molecules. This led Northrop and Simpson[104] to conclude that in the absence of an applied electric field, the number of electrons populating the conduction band is too small to account for the observed conductivity. Conductivity is due to the fact that there is greater population of electrons in the low lying bound states and there is a finite probability of ionization from the bound excited states due to the applied field. The mechanism is described by them as dissociation of the exciton in an applied field.

The exponential dependence of photoconductivity on temperature suggests that the charge carriers are formed through a molecular triplet so that for the actual separation for the uncoupled electrons, thermal energy is required. This explains the energy parameters E as activation energy having the same value 0.17 in both types of cell.

The question of the origin of the charge carriers has been somewhat controversial and much of the controversy derives from the effects of gases on the photocurrent. It has been observed by many[116] that the surface current is profoundly enhanced due to the presence of adsorbed gases such as O_2, SO_2, etc. This led Bree

[116] H. J. Zinzer, Z. Naturforsch. 11a, 306 (1956).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

and Lyons to suggest that oxygen serves as a catalyst in the formation of a triplet level which reacts to form the biradical intermediate AOO. When the AOO dissociates into ions, charges are liberated and the positive holes move through the lattice. This scheme of formation of charge carriers is, however, not properly equipped to explain the origin of bulk photocurrent established by Compton *et al.* who regard fluorescence and photoconductivity as processes competing with each other when an exciton ends its life. This is supported by experimental investigations of Zinzer[116] who found that prolonged α-ray bombardment reduces the fluorescence and the surface photocurrent and by those of Compton *et al.* who observe that neutron bombardment reduces both bulk and surface photocurrent as well as fluorescence. Compton *et al.* observe that the incorporation of tetracene into anthracene lattice also reduces the photocurrent in the 3800 system of anthracene, though Carswell and Lyons find that introduction of 10^{-4} M tetracene into anthracene increases the photocurrent, especially in the region of the second electronic transition (40 000 cm). Many workers[21,2] observed that incorporation of tetracene or acridine into the anthracene lattice suppresses or quenches the fluorescence. The quantum efficiency of photoconductivity of anthracene is 1. The former figure may mean that the range of charge carriers is small compared with the electrode separation which is suported by the observed lack of saturation in the photocurrent. A knowledge of the carrier mobility is required to convert detected currents into quantum efficiency of charge formation, and it remains to be seen whether there is quantum equivalence between absorption and photoconductivity.

7. SOME NEW RESULTS

The contents of the Secs. 3, 4, and 5 seem to indicate that, in most of the aromatic crystals investigated, the energy of excitation can migrate away from the place of excitation in the form of excitons, and that if the exciton is annihilated deep in such a crystal with the emission of fluorescence radiation, the emitted radiation may be absorbed in a photon cascade process by the crystal having intense absorption band in the region of emission. But the sections do not furnish any detailed information regarding the place of terminus of the energy transport, and it is not known how the final stage of the transfer is executed in a pure crystal, nor in an impure one. Polarization properties of crystal absorption and emission are expected to yield valuable information in this regard. If the crystal fluorescence is due to the reverse of photon absorption, fluorescence should give the same polarization ratio as absorption, provided both correspond to the same electronic transition. Experimental observations show that the polarization ratio of absorption (3800 A) and the fluorescence (corresponding to the same electronic transition) spectra are very different.

The polarization ratio for absorption of light in the c' direction of an anthracene crystal, when c' is perpendicular to the a and b crystallographic axes, has been measured by Craig and Hobbins[43] to be 1.8:1 and by Ganguly, Chaudhury, and Mukherjee[49] to be 1.5:1. The ratio of fluorescence intensity along the same directions, observed in the c' direction, has been determined by Ferguson and Schneider[117] to have the value 5:1 and the value obtained by Chaudhury[118] who corrected for all probable errors, is 4.3:1. Ganguly and Chaudhury[119] calculated on the basis of an oriented gas model that the polarization ratio (I_b/I_a) is given by the square of the cotangent of the angle which the active molecular axis (projected on the ab face) makes with b axis of the crystal. Craig and Hobbins[43] showed that the first-order perturbation methods applied to the exciton theory gives the same result. For an anthracene crystal it is 7.8:1 which is reduced to 6.5:1 on making allowance for torsional oscillations. Ferguson and Schneider argue that if the value 1.8:1 for absorption is accounted for, as was done by Craig,[120] by assuming a crystal induced mixing of the first two excited states of an anthracene molecule, the value of 4–5:1 for fluorescence, being nearer the exciton theory prediction, implies that the fluorescence is emitted from a molecule situated at a lattice dislocation where the symmetry elements of the lattice are not maintained. The different values for absorption and fluorescence I_b/I_a may be due to the fact that fluorescence is emitted from a molecule situated at a place where the principal directions of transition are rotated with regard to those of crystal absorption. Thus, these facts seem to establish the idea that an exciton in an anthracene crystal ends its life by being captured by a molecule situated at a dislocation from where the fluorescence is emitted.

When naphthacene molecules are present in an anthracene crystal that is excited by ultraviolet radiation, the naphthacene molecules may fluoresce by any one of the processes described in Sec. 1. If the photon cascade process is responsible, the decay time for fluorescence of anthracene would be different from the decay time of fluorescence of naphthacene in a crystal containing naphthacene as impurity. Unfortunately, there are hardly any useful data on decay times of fluorescence in mixed crystals, except those of Bose[74] which are not free from ambiguity. In this laboratory we have measured the optical densities of an anthracene crystal containing naphthacene molecules and calculated the effect of naphthacene absorption on anthracene fluorescence, according to the formula[118]

$$\frac{f_r}{f_0}=\frac{\mu_1 l}{\mu_1 l+\mu_2 l}\left\{\frac{1-e^{-(\mu_1+\mu_2)l}}{1-e^{-\mu_1 l}}\right\},$$

[117] J. Ferguson and W. G. Schneider, J. Chem. Phys. **25**, 780 (1956).
[118] N. K. Chaudhury, Z. Physik. **151**, 93–105 (1958).
[119] S. C. Ganguly and N. K. Chaudhury, Phys. Rev. **95**, 1148 (1954).
[120] D. P. Craig, J. Chem. Soc. 2302 (1955).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

where f_r is the backward fluorescence intensity of anthracene if there is absorption, f_0 is the same for no absorption; $\mu_1 t$ is the optical density of this crystal at the exciting wavelength λ_1, $\mu_2 t$ that at the emitted wavelength λ_2. In our measurements, $t=9.9\times10^{-3}$ cm, $\mu_1 t\sim10^3$ at $\lambda_1=3650$ A; $\mu_2 t\sim1$ at $\lambda_2=4898$ A where the absorption intensity of naphthacene molecules is the greatest in the fluorescence region of anthracene and $f_r/f_0\simeq1$. Thus absorption cannot be held responsible for transfer of energy from anthracene to naphthacene.

It has also been observed in this laboratory that when naphthacene molecules are incorporated within the anthracene lattice, not only the absorption and fluorescence spectra of naphthacene molecules show polarization, but also the polarization ratio of absorption spectra of these molecules remains unaltered from that of pure anthracene crystal. This means that the naphthacene molecules are embedded in an anthracene lattice as substitutes for anthracene molecules, and also that they are coupled to the anthracene lattice in the same way as anthracene molecules themselves. Thus it may be inferred that if the fluorescence of anthracene occurs at molecules situated in imperfect regions of crystal lattice, so it is for naphthacene fluorescence. The transfer of excitation energy from anthracene to naphthacene molecules is half complete at a concentration of $c=1.8\times10^{-5}$ mole naphthacene per mole anthracene.[28] If the naphthacene molecules occur at lattice dislocations at the same concentration as anthracene molecules, which is quite reasonable in view of the foregoing facts, the probability that a naphthacene molecule receives the excitation energy, and emits it in the form of fluorescence, should be at least 5×10^4 times as great as that of an anthracene molecule.

When the intensity ratio of the polarized fluorescence band at 4450 A of an anthracene crystal along the *b* axis to that along the *a* axis is plotted against different concentrations of naphthacene in the crystal, it is observed that the ratio at first gradually increases, attains a maximum value, then decreases with increase of concentration of naphthacene. This increase of intensity ratio with increase of concentration of naphthacene at low concentration region excludes absorption process. But the fall of the intensity ratio of fluorescence of anthracene at high concentration region of naphthacene is difficult to explain if the absorption process is prohibited altogether.

8. SUMMARY

Though our knowledge of energy transfer in aromatic crystals has increased considerably, it has not reached the state of maturity required for a definite formulation of a general rule. Many aspects of the problem remain almost untouched. Except for Bowen *et al.* and Birks *et al.*, very few workers have measured the effect of crystal size and shape on the fluorescence intensity, nor the efficiency of transfer of energy in these crystals. Because a quantum mechanical resonance transfer process is independent of crystal size and shape, the importance of such measurements in distinguishing the photon cascade from resonance transfer cannot be overemphasized. More and more careful measurements would be worthwhile. There exist few data on the efficiency of energy transfer from a host molecule to an impurity as a function of the exciting wavelength. It is also not known whether naphthacene fluoresces with the same efficiency by transferred excitation energy as by direct absorption, though in a recent paper Simpson[60] has shown that the intensity of naphthacene fluorescence excited by absorption of fluorescence of anthracene is smaller than that under exciton excitation. Because of complete quenching of the fluorescence of anthracene, it cannot be concluded whether the fluorescence efficiency is or is not the same in both the cases. Careful study of the results show that the agreements between the predictions of the exciton theory of the molecular organic crystal and the observed spectroscopic data are only over all and not in details. Again, cascade theory cannot explain many experimental results, though it gives satisfactory solutions to the problem of mechanism of energy transfer in pure crystal. This shows that either data of greater precision or some modification in the theories must be introduced. Measurements of absorption coefficients have not been made in many of these crystals, pure or impure, with the result that values of the absorption coefficients in solutions are quoted in support of a theory in solids. The importance of temperature in emission and absorption cannot be denied, but it has been almost completely ignored. More data are required on photoconductivity. Results obtained by different workers on the same crystals not only differ but are sometimes contradictory: how the mobility is influenced by various factors needs to be known for clearer understanding of the subject. It is expected that the study of temperature effect on the efficiency of fluorescence, coefficients of absorption, energy transfer, and polarization will reveal facts of far-reaching consequences. We have undertaken some of the studies in this laboratory.

Note added in proof.—After the review had been sent for publication, some new works having some bearing on the subject matter of the review appeared. They are briefly described below.

Electronic states of aromatic crystals (Sec. 3).—The near ultraviolet polarized absorption spectra of some aromatic crystals of pyrene have been investigated by Ferguson[121] at room temperature and at 77°K, of 1,3,5-trichlorobenzene by Schnepp[122] at 42°K, of hexachlorobenzene by Schnepp and Kopelman[123] at liquid nitrogen and liquid helium temperature, and of durene by Schnepp and McClure[124] at 20°K. Whereas in the case of pyrene, Davydov splitting could be detected as a separation of 18 cm^{-1} between the *a*- and *b*-polarized components, the *a* component being of higher energy in 1,3,5-trichlorobenzene, hexachlorobenzene, and durene

[121] J. Ferguson, J. Chem. Phys. **28**, 765 (1958).
[122] O. Schnepp, J. Chem. Phys. **30**, 863 (1959).
[123] O. Schnepp and R. Kopelman, J. Chem. Phys. **30**, 368 (1959).
[124] O. Schnepp and D. S. McClure, J. Chem. Phys. **30**, 874 (1959).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

crystals, no such splitting could be detected. If present, the splitting must not be greater than 5 cm^{-1} in the trichlorobenzene and hexachlorobenzene and 3 cm^{-1} in the durene crystal. In the case of the hexachlorobenzene crystal, the energy of separation between the two $K=0$ levels in the exciton band has been calculated to be 0.7 cm^{-1}.

Thus it appears that at least one of the low-lying excited electronic states of the pyrene crystal has the nature of exciton states, though in the case of the 1,3,5-trichlorobenzene, hexachlorobenzene, and durene crystals any such conclusion would be too hasty in view of the paucity of experimental data.

Photoconductivity of organic crystals (Sec. 6).—In a recent paper[125] Northrop and Simpson have published values of photocurrents of pure hydrocarbon crystals. They observed that the impurity molecules embedded in pure crystals which quench the fluorescence of pure substance reduce the photocurrent in the same ratio. Their other observations regarding dependence of photocurrent with light intensity, applied voltage, etc., are in agreement with previous publications. From these observations they propose that the interaction of two excitons is required to produce a single ionized molecule. Thus the production of charge carriers is explained. Compton *et al.*[126] studied photocurrent of anthracene crystals before and after neutron bombardment and concluded that conductivity was greatly reduced on bombardment. They further observed that dependence of photocurrent on wavelength and intensity of incident light remained unaltered after bombardment. Before bombardment the photocurrent was markedly nonohmic, after bombardment it was ohmic up to a field of 25 000 volts cm^{-1}. Before neutron bombardment the sample shows change of photocurrent when polarity of electrode is reversed, but this asymmetry disappears after irradiation. Kommandeur and Schneider[127] studied the photoconductivity in greater detail with very pure specimens of anthracene crystals and obtained results very different from previous ones. They observed that the maximum value of photocurrent corresponds to the minima of the absorption spectra. They also observed that the intensity dependence of photocurrent changes with wavelength, field direction, and even with magnitude of the applied field. These authors finally concluded that spectral response, voltage, and intensity dependence of photocurrent depend on the source and treatment of the crystals used, i.e., it depends on the density of imperfections of the crystals.

[125] D. C. Northrop and O. Simpson, Proc. Roy. Soc. (London) A244, 377 (1958).

[126] Compton, Schneider, and Waddington, J. Chem. Phys. 28, 741 (1958).

9. ACKNOWLEDGMENTS

We take this opportunity to express our sincere thanks to the Government of West Bengal for kindly extending a research grant to one of us (S.C.G.).

[127] J. Kommandeur and W. G. Schneider, J. Chem. Phys. 28, 582, 590 (1958).

REVIEWS OF MODERN PHYSICS VOLUME 31, NUMBER 4 OCTOBER, 1959

Irreversible Thermodynamics of Nonlinear Processes and Noise in Driven Systems*

WILLIAM BERNARD† ‡ AND HERBERT B. CALLEN

University of Pennsylvania, Philadelphia 4, Pennsylvania

1. INTRODUCTION

THIS paper reviews and extends the theory of irreversible thermodynamics. The irreversible behavior of a system driven by externally applied forces has been studied extensively, but attention has been focused primarily on the first-order term in the driven response. Here we also consider the higher-order terms in the driven response and the random fluctuations, or noise, occurring during an irreversible process. In addition to the well-known relations between the linear response and the equilibrium fluctuations, several new relations are proved involving the nonlinear response, the driven noise, and the equilibrium fluctuations.

The method of analysis is statistical mechanical and general, neither assuming a specific model nor postulating Markoffian behavior. The purposes of the analysis are thermodynamic; that is, to investigate interrelationships among macroscopically observable characteristics of systems undergoing irreversible processes. In this sense the aim should be clearly differentiated from those other approaches which might be characterized as kinetic or statistical mechanical rather than thermodynamic.

The most direct approach to the problem of irreversibility is the kinetic approach, in which a specific model is immediately introduced. The essential features of the model may be expressed in terms of molecular collision probabilities, giving rise to the Boltzmann equation, or to some similarly detailed kinetic equation. This is the standard method of "transport theory," and it is the method characteristic of the theory of the solid state.

A considerably more general approach is one which we term the irreversible statistical mechanical approach. The purpose there is to develop a general formalism, analogous to the partition sum algorithm of equilibrium statistical mechanics, which would provide a systematic recipe for the calculation of any macroscopically ob-

* This work was supported in part by the Office of Naval Research.

† Now at Research Division, Raytheon Company, Waltham, Massachusetts.

‡ Recipient of Philco Physics Fellowship, 1956–1958.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

servable characteristic of a system undergoing an irreversible process. No specific model is invoked; the aim is rather to provide a general formalism into which any particular model could be substituted to obtain explicit results. The irreversible statistical mechanical approach has not been completely successful as yet, but one type of partial result has been exploited widely. In this type of result the driven response of a system is obtained as a perturbation expansion in the applied forces. The various order response terms are typically expectation values of (multiple) commutators, taken with respect to the equilibrium system. It is, of course, hoped that a general algorithm for the computation of equilibrium commutator forms then will be developed to complete the general formalism.

The third approach is the thermodynamic approach which we adopt here. Although the statistical mechanical formalism is used to describe the motion, our purpose is not to compute either the response functions or the value of any quantities characterizing the equilibrium system. Our purpose is rather to explore the general interrelationships among different types of response functions and the equilibrium fluctuations, insisting, however, that the quantities so related each be macroscopically observable. Thus, for example, the equilibrium commutator forms in terms of which statistical mechanics expresses various response functions are not true observables of the equilibrium system. In order to give thermodynamic significance to statistical mechanical results, it is therefore necessary to re-express such quantities in terms of macroscopically observable symmetrized equilibrium forms, or anticommutators.

Three general classes of irreversible thermodynamic results have previously been obtained for first-order processes: (a) relationships between off-diagonal elements of the admittance—these are the Onsager reciprocity,[1,2] and its extension to non-Markoffian systems[3]; (b) the relationship between the first-order response and the second correlation moments of the equilibrium fluctuations—this is the so-called fluctuation-dissipation theorem[4,5]; (c) the relationship between the path distribution function for a driven system and the equilibrium fluctuations.[6–8]

The extensions of the theory which are developed here, and the general structure of irreversible thermodynamics, are summarized in the diagram in Fig. 1. The quantities appearing at the vertices of the diagrams denote the macroscopic observables, while the connecting lines indicate the existence of thermodynamic relationships. The arrowheads refer to the direction in

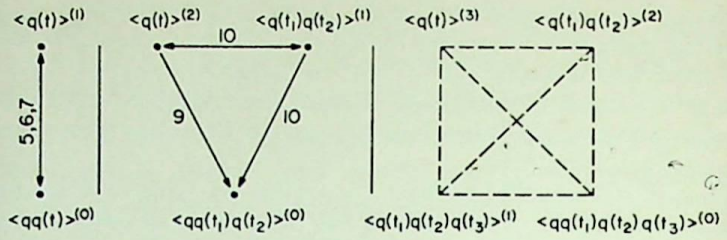


FIG. 1.

which specific relationships are developed in this paper, the numbers along the lines indicating the sections in which the various proofs appear.

The nature of the observables of interest can be made clear by the following considerations. Let $\langle q(t)\rangle$ denote the expectation value of the variable q at time t in a system driven by externally applied forces; that is, the driven response of the variable q. Further, let $\langle q(t)\rangle$ be expanded in powers of the applied forces. Then we label the first-order term in the response $\langle q(t)\rangle^{(1)}$, the second-order term $\langle q(t)\rangle^{(2)}$, etc. The zeroth-order term $\langle q(t)\rangle^{(0)} = \langle q\rangle^{(0)}$ is simply the average value of q in the equilibrium system. We similarly obtain a more detailed description of the driven system by introducing the expectation value $\langle q(t_1)q(t_2)\rangle$ of the product of the variable q at time t_1 with its value at time t_2; this is the second correlation moment of the random fluctuations, or noise, in the driven system. Again it is possible to expand $\langle q(t_1)q(t_2)\rangle$ in powers of the applied forces. We label the first-order term in the driven noise $\langle q(t_1)q(t_2)\rangle^{(1)}$, the second-order term $\langle q(t_1)q(t_2)\rangle^{(2)}$, and so on. The zeroth-order term $\langle q(t_1)q(t_2)\rangle^{(0)} = \langle qq(t_2-t_1)\rangle^{(0)}$ characterizes the spontaneous fluctuations in the equilibrium system. Similarly, it is possible to consider third- and high-order driven correlation moments.

There exists a definite hierarchy of irreversible thermodynamic relationships. The left-hand diagram represents the fluctuation-dissipation theorem, between the first-order response $\langle q(t)\rangle^{(1)}$ and the equilibrium second moment $\langle qq(t)\rangle^{(0)}$. The middle diagram indicates the triplet of relationships which exists among the second-order response $\langle q(t)\rangle^{(2)}$, the first-order term $\langle q(t_1)q(t_2)\rangle^{(1)}$ in the driven second moment (noise), and the third moment $\langle qq(t_1)q(t_2)\rangle^{(0)}$ of the spontaneous equilibrium fluctuations. The right-hand diagram indicates the cycle of interrelationships which may be presumed to exist among the next appropriate group of observables, although we do not consider this case explicitly.

In Secs. 2 to 4 the general statistical mechanical description of the time-evolution of a driven system is briefly reviewed. Sections 5 to 8 are devoted to a review of the existing first-order theory of irreversible thermodynamics. In Secs. 9 to 14 the general theory of irreversible thermodynamics is extended to the second-order response and the driven noise. Sections 13 and 14 are devoted to the irreversible thermodynamics of step-driven processes. In Secs. 15 and 16 the question of path distribution functions is considered, which may be regarded as the fundamental quantities of irreversible

[1] L. Onsager, Phys. Rev. **37**, 405 (1931); **38**, 2265 (1931).
[2] H. B. G. Casimir, Revs. Modern Phys. **17**, 343 (1945).
[3] Callen, Barasch, and Jackson, Phys. Rev. **88**, 1382 (1952).
[4] H. B. Callen and T. A. Welton, Phys. Rev. **83**, 34 (1951).
[5] H. B. Callen and R. F. Greene, Phys. Rev. **86**, 702 (1952); **88**, 1387 (1952).
[6] L. Onsager and S. Machlup, Phys. Rev. **91**, 1505 (1953); S. Machlup and L. Onsager, *ibid.* **91**, 1512 (1953).
[7] L. Tisza and I. Manning, Phys. Rev. **105**, 1695 (1956).
[8] H. B. Callen, Phys. Rev. **111**, 367 (1958).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

thermodynamics in the sense that all macroscopic quantities are derivable therefrom.

3. THE TIME EVOLUTION OF DRIVEN OPERATORS

We consider a system, of which the unperturbed Hamiltonian is $H^{(0)}$, in interaction with a number of external driving systems or signal generators. The Hamiltonian of the composite system may typically be represented by

$$H=H^{(0)}+\sum_i F_iQ_i+H_{sg}, \tag{1}$$

where $Q_i=Q_i(q,p)$ is a function of the coordinates q and momenta p of the system of interest, $F_i=F_i(q',p')$ is a function of the coordinates q' and momenta p' of the ith signal generator, and H_{sg} denotes the Hamiltonians of the signal generators.

Whereas the dissipative system possesses a large number of degrees of freedom and a quasi-continuous spectrum of energy eigenvalues, the signal generators have relatively few degrees of freedom and an extremely high degree of excitation. Thus, the coordinates of the signal generators, and consequently the $F_i(q',p')$, are essentially classical functions of the time. Ignoring the term H_{sg} as being irrelevant to the system of interest, the perturbed Hamiltonian assumes the form

$$H=H^{(0)}+\sum_i F_i(t)Q_i. \tag{2}$$

We adopt the interpretation that the Hermitian operators $Q_i(q,p)$ correspond to thermodynamic extensive parameters of the system of interest, while the $F_i(t)$ represent the conjugate intensive parameters imposed upon the system by the various signal generators. Thus, $Q_i(q,p)$ might be the operator corresponding to the position of a movable piston (volume), the total number of particles, or the magnetic moment of the system. The respective imposed intensive parameter would then be the pressure, the electrochemical potential, or the applied magnetic field.

If the system is in equilibrium, with the temperature T, before the forces are applied, the expectation value of an operator Q_i at time t is

$$\langle Q_i(t)\rangle=\text{Trace}\,\rho^{(0)}Q_i(t), \tag{3}$$

where $\rho^{(0)}$ is the initial (unperturbed) canonical density operator

$$\rho^{(0)}=\exp[-\beta H^{(0)}]/\text{Tr}\exp[-\beta H^{(0)}],\quad \beta=1/kT, \tag{4}$$

and where the Heisenberg operator $Q_i(t)$ is defined by

$$Q_i(t)=U^\dagger(t)Q_iU(t). \tag{5}$$

The unitary time-evolution operator $U(t)$ satisfies the Schrödinger equation

$$\dot{U}(t)=\frac{1}{i\hbar}H(t)U(t)=\frac{i}{i\hbar}[H^{(0)}+\sum_i F_i(t)Q_i]U(t). \tag{6}$$

The differential equation (6) is equivalent to the integral equation

$$U(t)=\exp\left[-i\frac{H^{(0)}t}{\hbar}\right]\left[1+\frac{1}{i\hbar}\sum_i\int_{-\infty}^{t}dt_1F_i(t_1)Q_i^{(0)}(t_1)\times\exp\left[i\frac{H^{(0)}t_1}{\hbar}\right]U(t_1)\right] \tag{7}$$

the correctness of which can be verified by differentiation. $Q_i^{(0)}(t)$ represents the unperturbed Heisenberg operator

$$Q_i^{(0)}(t)=\exp\left[i\frac{H^{(0)}t}{\hbar}\right]Q_i\exp\left[-i\frac{H^{(0)}t}{\hbar}\right]. \tag{8}$$

The iterative solution of Eq. (7) is

$$U(t)=\exp\left[-i\frac{H^{(0)}t}{\hbar}\right]\sum_{n=0}^{\infty}\left(-\frac{1}{i\hbar}\right)^n\sum_{ij\cdots k}\int_{-\infty}^{t}dt_1\times\int_{-\infty}^{t_1}dt_2\cdots\int_{-\infty}^{t_{n-1}}dt_nF_i(t_1)F_j(t_2)\cdots\times F_k(t_n)Q_i^{(0)}(t_1)Q_j^{(0)}(t_2)\cdots Q_k^{(0)}(t_n). \tag{9}$$

The perturbation expansion of $Q_i(t)$ follows from Eqs. (5) and (9). The zeroth-order term is simply the unperturbed operator $Q_i^{(0)}(t)$. The first-order term $Q_i^{(1)}(t)$ is

$$Q_i^{(1)}(t)=-\frac{1}{i\hbar}\sum_i\int_{-\infty}^{t}dt_1F_j(t_1)[Q_j^{(0)}(t_1),Q_i^{(0)}(t)]_-. \tag{10}$$

The second-order term $Q_i^{(2)}(t)$ is

$$Q_i^{(2)}(t)=\left(-\frac{1}{i\hbar}\right)^2\sum_{jk}\left\{-\int_{-\infty}^{t}dt_1F_j(t_1)\times\int_{-\infty}^{t}dt_2F_k(t_2)Q_j^{(0)}(t_1)Q_i^{(0)}(t)Q_k^{(0)}(t_2)+\int_{-\infty}^{t}dt_1F_j(t_1)\int_{-\infty}^{t_1}dt_2F_k(t_2)\times[Q_k^{(0)}(t_2)Q_j^{(0)}(t_1)Q_i^{(0)}(t)+Q_i^{(0)}(t)Q_j^{(0)}(t_1)Q_k^{(0)}(t_2)]\right\}. \tag{11}$$

In order to put the second-order term into a more suitable form, we decompose the t_2 integral of the first

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

term, so that

$$\int_{-\infty}^{t} dt_1 F_j(t_1) \int_{-\infty}^{t} dt_2 F_k(t_2) Q_j^{(0)}(t_1) Q_i^{(0)}(t) Q_k^{(0)}(t_2)$$

$$= \int_{-\infty}^{t} dt_1 F_j(t_1) \int_{-\infty}^{t_1} dt_2 F_k(t_2) Q_j^{(0)}(t_1) Q_i^{(0)}(t) Q_k^{(0)}(t_2)$$

$$+ \int_{-\infty}^{t} dt_1 F_j(t_1) \int_{t_1}^{t} dt_2 F_k(t_2) Q_j^{(0)}(t_1) Q_i^{(0)}(t) Q_k^{(0)}(t_2)$$

$$= \int_{-\infty}^{t} dt_1 F_j(t_1) \int_{-\infty}^{t_1} dt_2 F_k(t_2) Q_j^{(0)}(t_1) Q_i^{(0)}(t) Q_k^{(0)}(t_2)$$

$$+ \int_{-\infty}^{t} dt_1 F_j(t_1) \int_{-\infty}^{t_1} dt_2 F_k(t_2) Q_k^{(0)}(t_2) \times Q_i^{(0)}(t) Q_j^{(0)}(t_1). \quad (12)$$

In the last step we have inverted the order of integration in the second integral and interchanged the dummy indices i, j and the dummy times t_1, t_2. Using this result, Eq. (11) can be written in the form

$$Q_i^{(2)}(t) = \left(-\frac{1}{i\hbar}\right)^2 \sum_{jk} \int_{-\infty}^{t} dt_1 F_j(t_1) \int_{-\infty}^{t_1} dt_2 F_k(t_2) \times [Q_k^{(0)}(t_2), [Q_j^{(0)}(t_1), Q_i^{(0)}(t)]_-]_-. \quad (13)$$

Examination of Eqs. (10) and (13) clearly indicates the general form of the nth-order term $Q_i^{(n)}(t)$ in the driven Heisenberg operator $Q_i(t)$. Thus, for example, the third-order term is

$$Q_i^{(3)}(t) = \left(-\frac{1}{i\hbar}\right)^3 \sum_{jkl} \int_{-\infty}^{t} dt_1 F_j(t_1) \int_{-\infty}^{t_1} dt_2 F_k(t_2) \times \int_{-\infty}^{t_2} dt_3 F_l(t_3) [Q_l^{(0)}(t_3) \times [Q_k^{(0)}(t_2), [Q_j^{(0)}(t_1), Q_i^{(0)}(t)]_-]_-]_-. \quad (14)$$

This section follows the perturbation formulation given by R. Kubo.[9] However, because we are later concerned with Q operators which are intrinsically time-dependent, we choose to examine the time evolution of $Q_i(t)$ rather than of $\rho(t)$.

3. THE MACROSCOPIC RESPONSE AND DRIVEN CORRELATION MOMENTS

The first-order term in the driven response $\langle Q_i(t)\rangle$ of the thermodynamic variable corresponding to the operator Q_i is

$$\langle Q_i(t)\rangle^{(1)} = \mathrm{Tr}\, \rho^{(0)} Q_i^{(1)}(t) = -\frac{1}{i\hbar} \sum_j \int_{-\infty}^{t} dt_1 F_j(t_1) \times \langle [Q_j^{(0)}(t_1), Q_i^{(0)}(t)]_-\rangle^{(0)} \quad (15)$$

where the bracket $\langle\ \rangle^{(0)}$ denotes an expectation value with respect to the equilibrium ensemble.

Similarly, the second-order term in the response is

$$\langle Q_i(t)\rangle^{(2)} = \left(-\frac{1}{i\hbar}\right)^2 \sum_{jk} \int_{-\infty}^{t} dt_1 F_j(t_1) \int_{-\infty}^{t_1} dt_2 F_k(t_2) \times \langle [Q_k^{(0)}(t_2), [Q_j^{(0)}(t_1), Q_i^{(0)}(t)]_-]_-\rangle^{(0)}. \quad (16)$$

The form of the higher-order terms is clear from Eqs. (15) and (16).

The spontaneous fluctuations in an equilibrium ensemble are characterized by the second correlation moments in time between each pair of variables, although all higher moments are required as well for a complete description. The second equilibrium correlation moment is

$$\Psi_{ij}^{(0)}(t) = \langle \tfrac{1}{2}[Q_i, Q_j^{(0)}(t)]_+\rangle^{(0)} \quad (17)$$

where the bracket $[\ ,\ ]_+$ denotes a symmetrized operator product, or anticommutator.

A driven ensemble also exhibits fluctuations about its average motion, which in general differ from the equilibrium fluctuations. The fluctuations in a driven ensemble are characterized by

$$\langle \tfrac{1}{2}[Q_i(t), Q_j(t+\tau)]_+\rangle = \mathrm{Tr}\, \rho^{(0)} \tfrac{1}{2}[Q_i(t), Q_j(t+\tau)]_+. \quad (18)$$

Using the result (10) for $Q_i^{(1)}(t)$, the first-order term $\langle \tfrac{1}{2}[Q_i(t), Q_j(t+\tau)]_+\rangle^{(1)}$ in the driven second moment (18) becomes

$$\langle \tfrac{1}{2}[Q_i(t), Q_j(t+\tau)]_+\rangle^{(1)} = \tfrac{1}{2}\langle [Q_i^{(0)}(t), Q_j^{(1)}(t+\tau)]_+ + [Q_i^{(1)}(t), Q_j^{(0)}(t+\tau)]_+\rangle^{(0)}$$

$$= -\frac{1}{2i\hbar} \sum_k \left\{ \int_{-\infty}^{t+\tau} dt_1 F_k(t_1) \langle [Q_i^{(0)}(t), \times [Q_k^{(0)}(t_1), Q_j^{(0)}(t+\tau)]_-]_+\rangle^{(0)} + \int_{-\infty}^{t} dt_1 F_k(t_1) \langle [[Q_k^{(0)}(t_1), Q_i^{(0)}(t)]_- \times Q_j^{(0)}(t+\tau)]_+\rangle^{(0)} \right\}. \quad (19)$$

Upon decomposition of the first integral according to

$$\int_{-\infty}^{t+\tau} dt_1 = \int_{-\infty}^{t} dt_1 + \int_{t}^{t+\tau} dt_1,$$

[9] R. Kubo, J. Phys. Soc. Japan 12, 570 (1957).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

Eq. (19) becomes

$$\langle\tfrac{1}{2}[Q_i(t), Q_j(t+\tau)]_+\rangle^{(1)}$$
$$=-\frac{1}{2i\hbar}\sum_k\Big\{\int_{-\infty}^{t} dt_1F_k(t_1)\langle[Q_k^{(0)}(t_1),$$
$$\times[Q_i^{(0)}(t), Q_j^{(0)}(t+\tau)]_+]_-\rangle^{(0)}$$
$$+\int_{t}^{t+\tau} dt_1F_k(t_1)\langle[Q_i^{(0)}(t),$$
$$\times[Q_k^{(0)}(t_1), Q_j^{(0)}(t+\tau)]_-]_+\rangle^{(0)}\Big\}. \quad (20)$$

The higher-order terms in the perturbation expansion of the driven second moment can be written in analogous fashion, although the expressions involved become rapidly more complicated. The second-order term is

$$\langle\tfrac{1}{2}[Q_i(t), Q_j(t+\tau)]_+\rangle^{(2)}$$
$$=\frac{1}{2}\Big(-\frac{1}{i\hbar}\Big)^2\sum_{kl}\Big\{\int_{-\infty}^{t} dt_1F_k(t_1)\int_{-\infty}^{t_1} dt_2F_l(t_2)$$
$$\times\langle[Q_l^{(0)}(t_2), [Q_k^{(0)}(t_1),$$
$$\times[Q_i^{(0)}(t), Q_j^{(0)}(t+\tau)]_+]_-]_-\rangle$$
$$+\int_{-\infty}^{t} dt_1F_k(t_1)\int_{t}^{t+\tau} dt_2F_l(t_2)$$
$$\times\langle[Q_k^{(0)}(t_1), [Q_i^{(0)}(t),$$
$$\times[Q_l^{(0)}(t_2), Q_j^{(0)}(t+\tau)]_-]_+]_-\rangle^{(0)}\Big\}. \quad (21)$$

The nth-order term in the driven response can be written directly in terms of the $(n-1)$st-order term in the driven commutator $\langle[Q_j(t_1),Q_i(t)]_-\rangle^{(n-1)}$.

$$\langle Q_i(t)\rangle^{(n)}=\frac{1}{n}\Big(-\frac{1}{i\hbar}\Big)\sum_j\int_{-\infty}^{t} dt_1F_j(t_1)$$
$$\times\langle[Q_j(t_1),Q_i(t)]_-\rangle^{(n-1)}. \quad (22)$$

For $n=1$ this expression is identical to Eq. (15). For $n=2$ it can be obtained directly from Eq. (16) and the commutator analog of Eq. (19). For any order it is a direct consequence of the iterative nature of the perturbation expansion, and furnishes a clear picture of the essential structure of the motion.

4. STEP-DRIVEN PROCESSES

The preceding two sections were concerned with the motion of a driven system for which the forces are arbitrary functions of the time. In order to illustrate the formalism in a simple manner, we also consider step-driven processes in particular.

A step-driven process is defined as one for which the generalized forces in the distant past have increased slowly from zero to some constant value. This constant force remains applied until $t=0$, at which time it is suddenly removed and the system is allowed to relax into its equilibrium configuration. For a step-driven process the first- and second-order responses reduce to

$$\langle Q_i(t)\rangle^{(1)}=-\frac{1}{i\hbar}\sum_j F_j\int_{-\infty}^{0} dt_1\langle[Q_j^{(0)}(t_1),Q_i^{(0)}(t)]_-\rangle^{(0)}, \quad (23)$$

$$\langle Q_i(t)\rangle^{(2)}=\Big(-\frac{1}{i\hbar}\Big)^2\sum_{jk}F_jF_k\int_{-\infty}^{0} dt_1\int_{-\infty}^{t_1} dt_2$$
$$\times\langle[Q_k^{(0)}(t_2),[Q_j^{(0)}(t_1),Q_i^{(0)}(t)]_-]_-\rangle^{(0)}. \quad (24)$$

When the indicated time integrations are performed in Eqs. (23) and (24), the contributions from the infinite time limits pose certain difficulties, which are related to the approach to equilibrium. This matter is examined in Appendix A. It is also possible to approach the question of step-driven processes from the following alternate point of view, which circumvents these difficulties. We expect the applied forces F_j to bring the system to a new equilibrium configuration at $t=0$, characterized by a density operator $\rho(0)$ having the generalized canonical form appropriate to an ensemble in contact with a set of reservoirs with constant intensive parameters F_j.

$$\rho(0)=\exp\{-\beta[H^{(0)}+\textstyle\sum_j F_jQ_j]\}/$$
$$\operatorname{Tr}\exp\{-\beta[H^{(0)}+\textstyle\sum_j F_jQ_j]\}. \quad (25)$$

At $t=0$ the interaction with the external systems is removed. Since for $t>0$ the Hamiltonian is simply $H^{(0)}$, the response $\langle Q_i(t)\rangle$ during a step-driven process is given by

$$\langle Q_i(t)\rangle=\operatorname{Tr}\rho(0)Q_i^{(0)}(t)$$
$$=\operatorname{Tr}\exp\{-\beta[H^{(0)}+\textstyle\sum_j F_jQ_j]\}Q_i^{(0)}(t)/$$
$$\operatorname{Tr}\exp\{-\beta[H^{(0)}+\textstyle\sum_j F_jQ_j]\}. \quad (26)$$

In order to expand Eq. (26), we first perform the well-known expansion of the operator

$$A(\beta)\equiv\exp\{-\beta[H^{(0)}+\epsilon H^{(1)}]\},$$

where $H^{(1)}$ denotes the perturbation Hamiltonian. $A(\beta)$ satisfies the integral equation

$$A(\beta)=\exp[-\beta H^{(0)}]$$
$$\times\Big\{1-\epsilon\int_0^{\beta} d\lambda\exp[\lambda H^{(0)}]H^{(1)}A(\lambda)\Big\}, \quad (27)$$

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

the iterative solution of which is

$$A(\beta)=\exp[-\beta H^{(0)}]\sum_{n=0}^{\infty}(-\epsilon)^n\int_0^\beta d\lambda_1 H^{(1)}(-i\hbar\lambda_1)\times\int_0^{\lambda_1}d\lambda_2 H^{(1)}(-i\hbar\lambda_2)\cdots\times\int_0^{\lambda_{n-1}}d\lambda_n H^{(1)}(-i\hbar\lambda_n) \quad (28)$$

where $H^{(1)}(-i\hbar\lambda_1)=\exp[\lambda_1 H^{(0)}]H^{(1)}\exp[-\lambda_1 H^{(0)}]$.

Replacing $\epsilon H^{(1)}$ by the perturbation Hamiltonian $\sum_j F_j Q_j$ in Eq. (28) yields the quantity appearing in the numerator of Eq. (26).

$$\exp\{-\beta[H^{(0)}+\textstyle\sum_j F_j Q_j]\}=\exp[-\beta H^{(0)}]\sum_{n=0}^{\infty}(-1)^n\sum_{jk\cdots l}F_jF_k\cdots F_l\times\int_0^\beta d\lambda_1 Q_j^{(0)}(-i\hbar\lambda_1)\int_0^{\lambda_1}d\lambda_2 Q_k^{(0)}(-i\hbar\lambda_2)\cdots\times\int_0^{\lambda_{n-1}}d\lambda_n Q_l^{(0)}(-i\hbar\lambda_n). \quad (29)$$

The expansion of the denominator of Eq. (26) can be simplified by employing a technique, due to Nakajima,[10] which reduces all multiple temperature integrals by one order. Consider the trace of the operator $A(\beta)$ defined in the foregoing. Differentiating $\mathrm{Tr}\,A(\beta)=\mathrm{Tr}\exp\{-\beta[H^{(0)}+\epsilon H^{(1)}]\}$ with respect to ϵ, we obtain

$$\frac{\partial}{\partial\epsilon}\mathrm{Tr}\,A(\beta)=-\beta\,\mathrm{Tr}\exp\{-\beta[H^{(0)}+\epsilon H^{(1)}]\}H^{(1)}. \quad (30)$$

We now expand the quantity $\exp\{-\beta[H^{(0)}+\epsilon H^{(1)}]\}$ according to Eq. (27), integrate this expression with respect to ϵ, and substitute $\sum_j F_j Q_j$ for $\epsilon H^{(1)}$ to obtain the denominator of Eq. (25).

$$\mathrm{Tr}\exp\{-\beta[H^{(0)}+\textstyle\sum_j F_j Q_j]\}=\exp[-\beta H^{(0)}]\Big\{1+\beta\sum_{n=1}^{\infty}\frac{(-1)^n}{n}\sum_{j\cdots kl}F_j\cdots F_kF_l\times\int_0^\beta d\lambda_1\cdots\int_0^{\lambda_{n-2}}d\lambda_{n-1}Q_j^{(0)}(-i\hbar\lambda_1)\cdots\times Q_k^{(0)}(-i\hbar\lambda_{n-1})Q_l\Big\}. \quad (31)$$

We now substitute the expansions (29) and (31) into Eq. (26) and collect terms corresponding to each order in the perturbation. The results for the first few orders in the expansion of $Q_i(t)$ are found to be

$$\langle Q_i(t)\rangle^{(1)}=-\sum_j F_j\Big[\int_0^\beta d\lambda_1\langle Q_j^{(0)}(-i\hbar\lambda_1)Q_i^{(0)}(t)\rangle^{(0)}-\beta\langle Q_i\rangle^{(0)}\langle Q_j\rangle^{(0)}\Big], \quad (32)$$

$$\langle Q_i(t)\rangle^{(2)}=\sum_{jk}F_jF_k\Big[\int_0^\beta d\lambda_1\int_0^{\lambda_1}d\lambda_2\langle Q_j^{(0)}(-i\hbar\lambda_1)\times Q_k^{(0)}(-i\hbar\lambda_2)Q_i^{(0)}(t)\rangle^{(0)}-\frac{\beta}{2}\int_0^\beta d\lambda_1\langle Q_j^{(0)}(-i\hbar\lambda_1)Q_k\rangle^{(0)}\langle Q_i\rangle^{(0)}\Big]+\beta\sum_j F_j\langle Q_j\rangle^{(0)}\langle Q_i(t)\rangle^{(1)}, \quad (33)$$

$$\langle Q_i(t)\rangle^{(3)}=-\sum_{jkl}F_jF_kF_l\Big[\int_0^\beta d\lambda_1\int_0^{\lambda_1}d\lambda_2\int_0^{\lambda_2}d\lambda_3\times\langle Q_j^{(0)}(-i\hbar\lambda_1)Q_k^{(0)}(-i\hbar\lambda_2)\times Q_l^{(0)}(-i\hbar\lambda_3)Q_i^{(0)}(t)\rangle^{(0)}-\frac{\beta}{3}\int_0^\beta d\lambda_1\int_0^{\lambda_1}d\lambda_2\times\langle Q_j^{(0)}(-i\hbar\lambda_1)Q_k^{(0)}(-i\hbar\lambda_2)Q_l\rangle^{(0)}\langle Q_i\rangle^{(0)}\Big]+\beta\sum_j F_j\langle Q_j\rangle^{(0)}\langle Q_i(t)\rangle^{(2)}-\frac{\beta}{2}\sum_{jk}F_jF_k\times\int_0^\beta d\lambda_1\langle Q_j^{(0)}(-i\hbar\lambda_1)Q_k\rangle^{(0)}\langle Q_i(t)\rangle^{(1)}. \quad (34)$$

The classical forms of the foregoing equations are easily obtained and constitute a series of thermodynamic relationships. Letting $q_j(t)$ denote the time-dependent classical variable corresponding to the Heisenberg operator $Q_j(t)$ and replacing traces by integrals $\int d\Gamma$ over phase space, the classical step-driven response is, from Eq. (26),

$$\langle q_i(t)\rangle=\int d\Gamma\rho(t)q_i=\int d\Gamma\exp\{-\beta[H^{(0)}+\textstyle\sum_j F_jq_j]\}q_i(t)\Big/\times\int d\Gamma\exp\{-\beta[H^{(0)}+\textstyle\sum_j F_jq_j]\} \quad (35)$$

[10] S. Nakajima, *Advances in Physics* (Taylor and Francis, Ltd., London, 1955), Vol. 4, p. 363.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

or

$$\langle q_i(t)\rangle = \langle \exp(-\beta \textstyle\sum_j F_j q_j) q_i(t)\rangle^{(0)} / \times \langle \exp(-\beta \textstyle\sum_j F_j q_j)\rangle^{(0)}. \quad (36)$$

Since all quantities appearing in Eq. (36) are now classical functions, the perturbation expansion can be carried out in a straightforward way. Thus, the first few order terms in the classical step-driven response $\langle q_i(t)\rangle$ are

$$\langle q_i(t)\rangle^{(1)} = -\beta \sum_j F_j [\langle q_j q_i(t)\rangle^{(0)} - \langle q_i\rangle^{(0)}\langle q_j\rangle^{(0)}], \quad (37)$$

$$\langle q_i(t)\rangle^{(2)} = \tfrac{1}{2}\beta^2 \sum_{jk} F_j F_k [\langle q_j q_k q_i(t)\rangle^{(0)} - \langle q_j q_k\rangle^{(0)}\langle q_i\rangle^{(0)}] + \beta \sum_j F_j \langle q_j\rangle^{(0)}\langle q_i(t)\rangle^{(1)}, \quad (38)$$

$$\langle q_i(t)\rangle^{(3)} = -\tfrac{1}{6}\beta^3 \sum_{jkl} F_j F_k F_l [\langle q_j q_k q_l q_i(t)\rangle^{(0)} - \langle q_j q_k q_l\rangle^{(0)}\langle q_i\rangle^{(0)}] - \tfrac{1}{2}\beta^2 \sum_{jk} F_j F_k \langle q_j q_k\rangle^{(0)}\langle q_i(t)\rangle^{(1)} + \beta \sum_j F_j \langle q_j\rangle^{(0)}\langle q_i(t)\rangle^{(2)}. \quad (39)$$

Equation (36) is the first truly thermodynamic relationship we have developed up to this point. It expresses the step-driven response $\langle q_i(t)\rangle$ in terms of the quantity $\langle \exp(-\beta \sum_j F_j q_j) q_i(t)\rangle^{(0)}$, which characterizes the spontaneous equilibrium fluctuations in an operationally significant way. In particular,

$$\langle \exp(-\beta \textstyle\sum_j (F_j q_j) q_i(t)\rangle^{(0)}$$

represents the second equilibrium correlation moment between the quantity $\exp(-\beta \sum_j F_j q_j)$ at time zero and the quantity q_i at time t.

The first-order term (37) in the expansion of Eq. (36) will be recognized as a classical form of the so-called fluctuation-dissipation theorem,[4] which relates the first-order response $\langle q_i(t)\rangle^{(1)}$ to the equilibrium second correlation moment $\langle q_j q_i(t)\rangle^{(0)}$ between the variables q_j and $q_i(t)$. We discuss the quantum-mechanical form of the fluctuation-dissipation theorem in the following three sections, considering the step-driven case specifically in Sec. 7.

Similarly, Eqs. (38), (39), etc., relate the second- and higher-order terms in the classical step-driven response $\langle q_i(t)\rangle$ to appropriate higher equilibrium fluctuation moments. The quantum-mechanical form of these relationships is presented in Sec. 13.

5. EQUILIBRIUM FLUCTUATIONS AND THE FIRST-ORDER RESPONSE

The linear theory of irreversibility is reviewed in this and the two following sections, following quite closely the formulation of R. Kubo.[9] The pattern of this development suggests the method of extension to nonlinear processes, and yields a relation between commutators and anticommutators to which we make frequent reference.

It is convenient to characterize the first-order response by the aftereffect function $\phi_{ij}^{(1)}(t)$, which is the response $\langle Q_j(t)\rangle^{(1)}$ to a δ-function force F_i applied at $t=0$. That is, by definition,

$$\langle Q_j(t)\rangle^{(1)} = \sum_i \int_{-\infty}^{t} dt_1 F_i(t_1)\phi_{ij}^{(1)}(t-t_1). \quad (40)$$

Thus, writing the equilibrium commutator in Eq. (15) in the equivalent form $\langle [Q_i, Q_j^{(0)}(t-t_1)]_-\rangle^{(0)}$, we identify

$$\phi_{ij}^{(1)}(t) = -\frac{1}{i\hbar}\langle [Q_i, Q_j^{(0)}(t)]_-\rangle^{(0)}. \quad (41)$$

The aftereffect function $\phi_{ij}^{(1)}(t)$ exhibits significant symmetry properties with respect to reversal of the time t and of an applied magnetic vector potential $\mathbf{A}$. The classical quantities q_i are assumed to be even functions of the particle velocities; explicitly indicating the dependence on $\mathbf{A}$, the operators $Q_i(\mathbf{A})$ then satisfy the relationship

$$Q_i(-\mathbf{A}) = Q_i^*(\mathbf{A}). \quad (42)$$

The unperturbed Hamiltonian $H^{(0)}(\mathbf{A})$ and its eigenfunctions also satisfy Eq. (42).

Because the response $\langle Q_j(t)\rangle^{(1)}$ must itself be real, it follows that $\phi_{ij}^{(1)}(t)$ is real.

A second property of $\phi_{ij}^{(1)}(t)$ is

$$\phi_{ij}^{(1)}(-t) = -\phi_{ji}^{(1)}(t). \quad (43)$$

Introducing the transformation $t \rightarrow -t$ in Eq. (41), we have

$$\phi_{ij}^{(1)}(-t) = -\frac{1}{i\hbar}\langle [Q_i, Q_j^{(0)}(-t)]_-\rangle^{(0)}. \quad (44)$$

The t dependence can be transferred to the operator Q_i by performing a unitary transformation with

$$\exp\{\pm i[H^{(0)}t/\hbar]\}, \text{ proving Eq. (43).}$$

$\phi_{ij}^{(1)}(t)$ is odd under reversal of time and magnetic field.

$$\phi_{ij}^{(1)}(-t; -\mathbf{A}) = -\phi_{ij}^{(1)}(t; \mathbf{A}). \quad (45)$$

According to Eq. (42) $Q_j^{(0)}(-t; -\mathbf{A}) = Q_j^{(0)*}(t; \mathbf{A})$, similarly $\rho^{(0)}(-\mathbf{A}) = \rho^{(0)*}(\mathbf{A})$, so that $\phi_{ij}^{(1)}(-t; -\mathbf{A}) = -\phi_{ij}^{(1)*}(t; \mathbf{A})$. Invoking the reality of $\phi_{ij}^{(1)}(t; \mathbf{A})$ yields the property (45).

Properties (43) and (45) also imply

$$\phi_{ij}^{(1)}(t; -\mathbf{A}) = \phi_{ji}^{(1)}(t; \mathbf{A}). \quad (46)$$

We now establish the fundamental relationship which exists between the equilibrium commutator

$$\phi_{ij}^{(1)}(t) = -(1/i\hbar)\langle [Q_i, Q_j^{(0)}(t)]_-\rangle^{(0)}$$

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

characterizing the first-order response, and the equilibrium anticommutator $\Psi_{ij}^{(0)}(t)=\langle\frac{1}{2}[Q_i,Q_j^{(0)}(t)]_+\rangle^{(0)}$, characterizing the spontaneous equilibrium fluctuations. The motivation for so doing is that the latter quantity is a true macroscopic observable of the equilibrium system, while the former is not.

Consider the equilibrium second moment $\Psi_{ij}^{(0)}(t)$. Since the analysis is carried out in the spectral representation, it is convenient to define the operators Q_i such that $\Psi_{ij}^{(0)}(t)$ has no constant component, thus avoiding the attendant δ-function singularity appearing in its Fourier transform. As discussed in Appendix A, the time-independent portion of $\Psi_{ij}^{(0)}(t)$ is

$$\lim_{t\to\infty}\Psi_{ij}^{(0)}(t)=\lim_{t\to\infty}\langle\tfrac{1}{2}[Q_i,Q_j^{(0)}(t)]_+\rangle^{(0)}$$
$$=\langle Q_i\rangle^{(0)}\langle Q_j\rangle^{(0)}. \quad (47)$$

Consequently, we assume that the operators Q_i are defined such that $\langle Q_i\rangle^{(0)}=0$.

$\Psi_{ij}^{(0)}(t)$ can be written

$$\Psi_{ij}^{(0)}(t)=\tfrac{1}{2}\langle Q_iQ_j^{(0)}(t)+Q_j^{(0)}(t)Q_i\rangle^{(0)}$$
$$=\tfrac{1}{2}\langle Q_iQ_j^{(0)}(t)+Q_i^{(0)}(-i\hbar\beta)Q_j^{(0)}(t)\rangle^{(0)}. \quad (48)$$

The second term on the right has been obtained by inserting $\exp[\pm\beta H^{(0)}]$ in front of Q_i and cyclically permuting the operators in the trace. That is,

$$\langle Q_j^{(0)}(t)Q_i\rangle^{(0)}=\mathrm{Tr}\,\rho^{(0)}Q_j^{(0)}(t)\exp[\pm\beta H^{(0)}Q_i$$
$$=\mathrm{Tr}\,\rho^{(0)}\exp[\beta H^{(0)}Q_i\exp[-\beta H^{(0)}]Q_j^{(0)}(t), \quad (49)$$

and invoking the definition (8) for $Q_i^{(0)}(-i\hbar\beta)$ gives Eq. (48). We decompose Eq. (48) into a double summation over matrix elements in the unperturbed energy representation

$$\Psi_{ij}^{(0)}(t)=\tfrac{1}{2}\sum_{lm}\rho(E_l)\{1+\exp[\beta(E_l-E_m)]\}$$
$$\times\langle E_l|Q_i|E_m\rangle\langle E_m|Q_j|E_l\rangle\exp\left[i\frac{(E_m-E_l)t}{\hbar}\right] \quad (50)$$

where $\rho(E_l)=e^{-\beta E_l}/\sum_l e^{-\beta E_l}$, and $\langle E_l|Q_i|E_m\rangle$ is the matrix element of Q_i between the eigenstates of $H^{(0)}$ having the eigenvalues E_l and E_m. In virtue of the quasi-continuous spectrum of energy eigenvalues, the double summation appearing in Eq. (50) can be replaced by a double integration over energy eigenvalues.

$$\Psi_{ij}^{(0)}(t)=\frac{1}{2}\int_{-\infty}^{\infty}dE_l\int_{-\infty}^{\infty}dE_m\rho(E_l)\eta(E_l)\eta(E_m)$$
$$\times\{1+\exp[\beta(E_l-E_m)]\}\langle E_l|Q_i|E_m\rangle$$
$$\times\langle E_m|Q_j|E_l\rangle\exp\left[i\frac{(E_m-E_l)t}{\hbar}\right] \quad (51)$$

where $\eta(E_l)$ is the energy density-of-states function.

We obtain the Fourier transform $G_{ij}^{(0)}(\omega)$ of $\Psi_{ij}^{(0)}(t)$ by introducing

$$E_m=E_l+\hbar\omega. \quad (52)$$

Thus Eq. (51) becomes

$$\Psi_{ij}^{(0)}(t)=\langle\tfrac{1}{2}[Q_i,Q_j^{(0)}(t)]_+\rangle^{(0)}$$
$$=\frac{1}{(2\pi)^{\frac{1}{2}}}\int_{-\infty}^{\infty}d\omega e^{i\omega t}G_{ij}^{(0)}(\omega) \quad (53)$$

where

$$G_{ij}^{(0)}(\omega)=\left(\frac{\pi}{2}\right)^{\frac{1}{2}}\hbar[1+\exp(-\hbar\beta\omega)]\int_{-\infty}^{\infty}dE\rho(E)$$
$$\times\eta(E)\eta(E+\hbar\omega)\langle E|Q_i|E+\hbar\omega\rangle$$
$$\times\langle E+\hbar\omega|Q_j|E\rangle; \quad (54)$$

$G_{ij}^{(0)}(\omega)$ is the spectrum of the spontaneous equilibrium fluctuations.

We obtain the Fourier transform of the aftereffect function $\phi_{ij}^{(1)}(t)$ in an analogous way. Equation (41) can be rewritten

$$\phi_{ij}^{(1)}(t)=-\frac{1}{i\hbar}\langle Q_iQ_j^{(0)}(t)-Q_j^{(0)}(t)Q_i\rangle^{(0)}$$
$$=-\frac{1}{i\hbar}\langle Q_iQ_j^{(0)}(t)-Q_i^{(0)}(-i\hbar\beta)Q_j^{(0)}(t)\rangle^{(0)}. \quad (55)$$

Decomposing the equilibrium expectation value into a double integral over matrix elements in the unperturbed energy representation and introducing the transformation (52), we obtain the result

$$\phi_{ij}^{(1)}(t)=\frac{1}{(2\pi)^{\frac{1}{2}}}\int_{-\infty}^{\infty}d\omega e^{i\omega t}L_{ij}^{(1)}(\omega) \quad (56)$$

where

$$L_{ij}^{(1)}(\omega)=(2\pi)^{\frac{1}{2}}i[1-\exp(-\hbar\beta\omega)]\int_{-\infty}^{\infty}dE\rho(E)\eta(E)$$
$$\times\eta(E+\hbar\omega)\langle E|Q_i|E+\hbar\omega\rangle\langle E+\hbar\omega|Q_j|E\rangle \quad (57)$$

is clearly the Fourier transform of $\phi_{ij}^{(1)}(t)$.

Comparison of Eqs. (54) and (57) shows that the Fourier transform $G_{ij}^{(0)}(\omega)$ of the equilibrium second moment $\Psi_{ij}^{(0)}(t)$ is related to the Fourier transform $L_{ij}^{(1)}(\omega)$ of the aftereffect function $\phi_{ij}^{(1)}(t)$ according to

$$i\omega G_{ij}^{(0)}(\omega)=E^{(1)}(\omega;\beta)L_{ij}^{(1)}(\omega) \quad (58)$$

where

$$E^{(1)}(\omega;\beta)=\frac{\hbar\omega}{2}\coth\frac{\hbar\beta\omega}{2}\underset{\beta\to0}{\longrightarrow}\frac{1}{\beta}. \quad (59)$$

The universal function $E^{(1)}(\omega;\beta)$ is uniquely quantum-mechanical in origin and corresponds to a slight

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

smearing out of the microscopic contributions to the macroscopic response at extremely high frequencies ($\gtrsim 10^{12}$ cps at room temperature). In the classical limit $\beta \to 0$, $E^{(1)}(\omega;\beta) \to (1/\beta)$ as indicated.

Equation (58) is the spectral statement of the fundamental relationship which exists between the first-order response and the spontaneous equilibrium fluctuations. It provides directly the basis for the fluctuation-dissipation theorem, several alternate forms of which have been developed.[3,4,9] We discuss these in the following two sections.

The result (58) has also been obtained by Kubo[9] using function-theoretical arguments rather than the matrix approach employed here.

6. THE ADMITTANCE AND THE FLUCTUATION-DISSIPATION THEOREM

We rephrase the results of the preceding section in the familiar terms afforded by the admittance matrix. We define $\alpha_j(\omega)$ and $\gamma_i(\omega)$ as the Fourier transforms of the first-order "current" and force, respectively,

$$\frac{d}{dt}\langle Q_j(t)\rangle^{(1)} = \langle \dot{Q}_j(t)\rangle^{(1)} = \frac{1}{(2\pi)^{\frac{1}{2}}}\int_{-\infty}^{\infty} d\omega e^{i\omega t}\alpha_j(\omega), \quad (60)$$

$$F_i(t) = \frac{1}{(2\pi)^{\frac{1}{2}}}\int_{-\infty}^{\infty} d\omega e^{i\omega t}\gamma_i(\omega). \quad (61)$$

We further define the admittance matrix elements $Y_{ij}(\omega)$ by

$$\alpha_j(\omega) = \sum_i \gamma_i(\omega) Y_{ij}(\omega) \quad (62)$$

whence, from the definition (40), it follows that

$$Y_{ij}(\omega) = i\omega\int_0^{\infty} dt e^{-i\omega t}\phi_{ij}^{(1)}(t). \quad (63)$$

By Eq. (63) the symmetry properties appropriate to $Y_{ij}(\omega)$ follow immediately from the symmetry properties of $\phi_{ij}^{(1)}(t)$. The reality of $\phi_{ij}^{(1)}(t)$ implies that the real and imaginary parts of $Y_{ij}(\omega)$ are even and odd, respectively, under the transformation $\omega \to -\omega$.

$$\left.\begin{matrix}\mathrm{Re}\\ \mathrm{Im}\end{matrix}\right\} Y_{ij}(-\omega) = \left\{\begin{matrix}+\mathrm{Re}\\ -\mathrm{Im}\end{matrix}\right\} Y_{ij}(\omega). \quad (64)$$

The symmetry property (46) of $\phi_{ij}^{(1)}(t)$ with respect to reversal of the applied magnetic field implies a corresponding symmetry of $Y_{ij}(\omega)$.

$$Y_{ij}(\omega; -\mathbf{A}) = Y_{ji}(\omega; \mathbf{A}). \quad (65)$$

Equation (65) represents the extension of the original Onsager reciprocity[1,2] to all frequency components of the admittance matrix elements.

We now rewrite the spectral relationship between the first order response and the equilibrium fluctuations in terms of the $Y_{ij}(\omega)$ by decomposing Eq. (58) into its symmetric and antisymmetric parts with respect to the indices ij.

$${}^{(s)(a)}G_{ij}^{(0)}(\omega) = \frac{E^{(1)}(\omega;\beta)}{i\omega}\tfrac{1}{2}[L_{ij}^{(1)}(\omega) \pm L_{ji}^{(1)}(\omega)] \quad (66)$$

where the superscripts (s) and (a) denote the symmetric and antisymmetric parts, respectively. However, using the time-reversal symmetry property (43) of $\phi_{ij}^{(1)}(t)$, it follows from Eq. (63) for $Y_{ij}(\omega)$ and Eq. (58) for $L_{ij}^{(1)}(\omega)$ that these quantities are related according to

$$\tfrac{1}{2}[L_{ij}^{(1)}(\omega) + L_{ji}^{(1)}(\omega)] = -\frac{2i}{(2\pi)^{\frac{1}{2}}}\frac{\mathrm{Re}^{(s)}Y_{ij}(\omega)}{\omega}, \quad (67)$$

$$\tfrac{1}{2}[L_{ij}^{(1)}(\omega) - L_{ji}^{(1)}(\omega)] = \frac{2}{(2\pi)^{\frac{1}{2}}}\frac{\mathrm{Im}^{(a)}Y_{ij}(\omega)}{\omega}. \quad (68)$$

$\mathrm{Re}^{(s)}Y_{ij}(\omega)$ is the real (symmetric) part of $Y_{ij}(\omega)$ and $\mathrm{Im}^{(a)}Y_{ij}(\omega)$ the imaginary (antisymmetric part). We note from Eq. (65) that $\mathrm{Re}^{(s)}Y_{ij}(\omega)$ and $\mathrm{Im}^{(a)}Y_{ij}(\omega)$ are even and odd, respectively, with respect to reversal of the vector potential $\mathbf{A}$.

Substituting the relations (67) and (68) into Eq. (66), we obtain the results

$${}^{(s)}G_{ij}^{(0)}(\omega) = -\left(\frac{2}{\pi}\right)^{\frac{1}{2}}E^{(1)}(\omega;\beta)\frac{\mathrm{Re}^{(s)}Y_{ij}(\omega)}{\omega^2}, \quad (69)$$

$${}^{(a)}G_{ij}^{(0)}(\omega) = -i\left(\frac{2}{\pi}\right)^{\frac{1}{2}}E^{(1)}(\omega;\beta)\frac{\mathrm{Im}^{(a)}Y_{ij}(\omega)}{\omega^2}. \quad (70)$$

Since $G_{ij}^{(0)}(\omega)$ is the spectrum of the equilibrium second moment $\langle\frac{1}{2}[Q_i,Q_j^{(0)}(t)]_+\rangle^{(0)}$, the Fourier transforms of Eqs. (69) and (70) are

$${}^{(s)}\langle\tfrac{1}{2}[Q_i,Q_j^{(0)}(t)]_+\rangle^{(0)} = -\frac{2}{\pi}\int_0^{\infty} d\omega\cos\omega t E^{(1)}(\omega;\beta)\frac{\mathrm{Re}^{(s)}Y_{ij}(\omega)}{\omega^2}, \quad (71)$$

$${}^{(a)}\langle\tfrac{1}{2}[Q_i,Q_j^{(0)}(t)]_+\rangle^{(0)} = \frac{2}{\pi}\int_0^{\infty} d\omega\sin\omega t E^{(1)}(\omega;\beta)\frac{\mathrm{Im}^{(a)}Y_{ij}(\omega)}{\omega^2}. \quad (72)$$

The symmetric part ${}^{(s)}\langle\frac{1}{2}[Q_i,Q_j^{(0)}(t)]_+\rangle^{(0)}$ of the equilibrium second moment with respect to ij is even with respect to reversal of the vector potential $\mathbf{A}$, while the antisymmetric part ${}^{(a)}\langle\frac{1}{2}[Q_i,Q_j^{(0)}(t)]_+\rangle^{(0)}$ is an odd function of $\mathbf{A}$. Further, ${}^{(a)}\langle\frac{1}{2}[Q_i,Q_j^{(0)}(t)]_+\rangle^{(0)}$ vanishes in the absence of an applied magnetic field.

Equations (69) and (70) or (71) and (72) constitute the familiar spectral statement of the fluctuation-dissipation theorem.[4] In the classical limit $\beta \to 0$ [see Eq. (59)], Eqs. (70) and (71) reduce to the familiar

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

Nyquist forms

$$^{(s)}\langle q_i q_j(t)\rangle^{(0)} = -\frac{2}{\pi\beta}\int_0^\infty d\omega \cos\omega t \frac{\mathrm{Re}^{(s)}Y_{ij}(\omega)}{\omega^2}, \quad (73)$$

$$^{(a)}\langle q_i q_j(t)\rangle^{(0)} = \frac{2}{\pi\beta}\int_0^\infty d\omega \sin\omega t \frac{\mathrm{Im}^{(a)}Y_{ij}(\omega)}{\omega^2}. \quad (74)$$

If we put $t=0$ in the Nyquist relation (73), the left-hand member represents the total noise intensity. It is of interest to note that this form of the equation for $t=0$ also follows from the Kramers-Krönig or dispersion relations, together with the results of equilibrium fluctuation theory. The Kramers-Krönig formulas relate the real and imaginary parts of the admittance matrix element $Y_{ij}(\omega)$ in consequence of the general requirement of causality. In Appendix B we consider this connection between the dispersion relations and the spectral form of the fluctuation-dissipation theorem.

It is sometimes convenient to characterize the spontaneous equilibrium fluctuations in terms of a set of hypothetical intensive quantities F_i rather than the extensive quantities Q_i. These hypothetical forces are associated with the fluctuating Q_i in the same formal way as real forces are associated with the average first-order driven response. That is, by analogy with Eq. (62), the fluctuating force is so defined that the product of its Fourier transform with $Y(\omega)/i\omega$ yields the Fourier transform of the fluctuating extensive parameter.

Consider a one-dimensional system with a single force $F(t)$ and corresponding operator Q. The spectrum $\mathcal{G}^{(0)}(\omega)$ of the second moment $\langle FF(t)\rangle^{(0)}$ of the equilibrium force fluctuation is defined by

$$\langle FF(t)\rangle^{(0)} = \frac{1}{(2\pi)^{\frac{1}{2}}}\int_{-\infty}^{\infty} d\omega e^{i\omega t}\mathcal{G}^{(0)}(\omega). \quad (75)$$

Using Eq. (62), together with Eq. (69) for $G^{(0)}(\omega)$, we obtain

$$\mathcal{G}^{(0)}(\omega) = \frac{\omega^2 G^{(0)}(\omega)}{|Y(\omega)|^2} = -\left(\frac{2}{\pi}\right)^{\frac{1}{2}} E^{(1)}(\omega;\beta)\frac{\mathrm{Re}\, Y(\omega)}{|Y(\omega)|^2}$$

$$= -\left(\frac{2}{\pi}\right)^{\frac{1}{2}} E^{(1)}(\omega;\beta)\,\mathrm{Re}\, Z(\omega) \quad (76)$$

where $\mathrm{Re}\, Z(\omega) = \mathrm{Re}\, Y(\omega)/|Y(\omega)|^2$ is the real, or dissipative, part of the complex impedance function. Thus, $\langle FF(t)\rangle^{(0)}$ is given by

$$\langle FF(t)\rangle^{(0)} = -\frac{2}{\pi}\int_0^\infty d\omega \cos\omega t E^{(1)}(\omega;\beta)\mathrm{Re}\, Z(\omega) \quad (77)$$

which corresponds to the one-dimensional form of the fluctuation-disspation theorem (71) for

$$\langle \tfrac{1}{2}[Q, Q^{(0)}(t)]_+\rangle^{(0)}.$$

Similarly, it is shown that the symmetric and antisymmetric parts of the second moment $\langle F_i F_j(t)\rangle^{(0)}$ of the equilibrium force fluctuations for a multidimensional system are given by

$$^{(s)}\langle F_i F_j(t)\rangle^{(0)} = -\frac{2}{\pi}\int_0^\infty d\omega \cos\omega t E^{(1)}(\omega;\beta) \times \mathrm{Re}^{(s)}Z_{ij}(\omega), \quad (78)$$

$$^{(a)}\langle F_i F_j(t)\rangle^{(0)} = \frac{2}{\pi}\int_0^\infty d\omega \sin\omega t E^{(1)}(\omega;\beta) \times \mathrm{Im}^{(a)}Z_{ij}(\omega). \quad (79)$$

The symmetry properties of the elements $Z_{ij}(\omega)$ of the complex impedance matrix are identical to those of $Y_{ij}(\omega)$.

The fluctuation-disspation theorem of Eqs. (71) and (72) or (78) and (79), which establishes a quantitative relationship between a dissipative process and appropriate equilibrium fluctuations, can be given the following intuitive interpretation. A dissipative process can be conveniently considered to involve the interaction between the dissipative system and a source system or signal generator. As mentioned at the beginning of Sec. 2, the dissipative system is characterized by a large number of degrees of freedom and is capable of absorbing energy when acted upon by an imposed force. In equilibrium it exhibits random fluctuations of its variables.

The source system, on the other hand, which provides the imposed forces and delivers energy to the dissipative system, is characterized by relatively few degrees of freedom and a high degree of excitation. Examples of such systems might be a classical pendulum or polyatomic molecule. When isolated from the dissipative system and given some internal energy, the source system may be regarded as having a sort of internal coherence.

If the source system is now connected to the dissipative system, this internal coherence is destroyed, the periodic motion vanishes, and the energy is sapped away, until finally the source system is left with only the random disordered energy $1/\beta$ characteristic of thermal equilibrium. This loss of coherence within the source system may be regarded as being caused by the random fluctuations generated by the dissipative system and acting back upon the source system itself. The disspation therefore appears as the macroscopic consequence of the disordering effect of the random equilibrium fluctuations, and, as such, is necessarily quantitatively related to the fluctuations.

An interesting analogy is furnished by the historical development of the theory of spontaneous radiation from excited atoms. After the initial development of quantum mechanics, it was found impossible to compute the spontaneous transition probabilities for an isolated excited atom, and this dissipative process appeared to be outside the existing structure of dynamics. With the

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

advent of quantum electrodynamics, however, the dissipation could be computed, and it was found that the spontaneous transitions could be consistently considered to be induced by the random fluctuations of the electromagnetic field in the vacuum. In this case, the excited atom plays the role of the source system, and the vacuum plays the role of the dissipative system.

7. THE FIRST-ORDER RESPONSE—TEMPORAL REPRESENTATION

A particularly useful temporal form of the fluctuation-dissipation theorem, due to R. Kubo,[9] is obtained by taking the Fourier transform of the basic spectral relationship (58) between $L_{ij}^{(1)}(\omega)$ and $G_{ij}^{(0)}(\omega)$. According to Eq. (56) this yields the aftereffect function $\phi_{ij}^{(1)}(t)$ in the form

$$\phi_{ij}^{(1)}(t)=\frac{1}{(2\pi)^{\frac{1}{2}}}\int_{-\infty}^{\infty}d\omega e^{i\omega t}\frac{1}{E^{(1)}(\omega;\beta)}i\omega G_{ij}^{(0)}(\omega). \quad (80)$$

Letting

$$1/E^{(1)}(\omega;\beta)=\int_{-\infty}^{\infty}dt e^{-i\omega t}\Gamma(t)$$

and noting from Eq. (53) that

$$i\omega G_{ij}^{(0)}(\omega)=[1/(2\pi)^{\frac{1}{2}}]\int_{-\infty}^{\infty}dt e^{-i\omega t}\langle\tfrac{1}{2}[Q_i,\dot{Q}_j^{(0)}(t)]_+\rangle^{(0)},$$

Eq. (80) becomes

$$\phi_{ij}^{(1)}(t)=\frac{1}{2\pi}\int_{-\infty}^{\infty}d\omega\int_{-\infty}^{\infty}dt'\Gamma(t')\times\int_{-\infty}^{\infty}dt''\langle\tfrac{1}{2}[Q_i,\dot{Q}_j^{(0)}(t'')]_+\rangle^{(0)}e^{i\omega(t-t'-t'')}. \quad (81)$$

Invoking the δ-function property of

$$\int_{-\infty}^{\infty}d\omega e^{i\omega(t-t'-t'')},$$

we obtain the result

$$\phi_{ij}^{(1)}(t)=-\frac{1}{i\hbar}\langle[Q_i,Q_j^{(0)}(t)]_-\rangle^{(0)}=\int_{-\infty}^{\infty}dt'\Gamma(t-t')\langle\tfrac{1}{2}[Q_i,\dot{Q}_j^{(0)}(t')]_+\rangle^{(0)} \quad (82)$$

where

$$\Gamma(t)=\frac{1}{2\pi}\int_{-\infty}^{\infty}d\omega e^{i\omega t}\frac{1}{E^{(1)}(\omega;\beta)}=\frac{2}{\pi\hbar}\ln\coth\frac{\pi|t|}{2\hbar\beta}\underset{\beta\to 0}{\longrightarrow}\beta\delta(t). \quad (83)$$

The evaluation of the function $\Gamma(t)$ has previously been carried out by R. Kubo,[9] and is presented in Appendix C.

Equation (82) constitutes the temporal statement of the fluctuation-dissipation theorem. Although equivalent to the spectral form of Eq. (58), it presents the basic relationship between equilibrium commutators and anticommutators in a more explicit fashion.

The result (82) can of course be substituted back into Eq. (40) for $\langle Q_j(t)\rangle^{(1)}$ to yield the first-order response directly in terms of the equilibrium correlation moment.

$$\langle Q_j(t)\rangle^{(1)}=-\sum_i\int_{\infty}^{t}dt_1F_i(t_1)\int_{-\infty}^{\infty}dt_1'\Gamma(t_1-t_1')\times\langle\tfrac{1}{2}[\dot{Q}_i^{(0)}(t_1'),Q_j^{(0)}(t)]_+\rangle^{(0)}. \quad (84)$$

In obtaining (84) from (40) and (82), we have introduced the transformation $t_1'=t-t'$ and made use of the identity

$$\langle\tfrac{1}{2}[Q_i,\dot{Q}_j^{(0)}(t_1-t_1')]_+\rangle^{(0)}=-\langle\tfrac{1}{2}[\dot{Q}_i^{(0)}(t_1'),Q_j^{(0)}(t)]_+\rangle^{(0)}.$$

In the classical limit $\beta\to 0$, Eq. (84), in virtue of the δ-function property of $\Gamma(t)$, reduces to

$$\langle q_j(t)\rangle^{(1)}=-\beta\sum_i\int_{-\infty}^{t}dt_1F_i(t_1)\langle\dot{q}_i(t_1)q_j(t)\rangle^{(0)}. \quad (85)$$

The first-order response during a step-driven process can be obtained directly from Eq. (84) by introducing the step-function forces defined in Sec. 4.

$$\langle Q_j(t)\rangle^{(1)}=-\sum_i F_i\int_{-\infty}^{0}dt_1\int_{-\infty}^{\infty}dt_1'\Gamma(t_1-t_1')\times\langle\tfrac{1}{2}[\dot{Q}_i^{(0)}(t_1'),Q_j^{(0)}(t)]_+\rangle^{(0)}. \quad (86)$$

Integrating by parts and putting $\langle Q_i\rangle^{(0)}=0$ gives

$$\langle Q_j(t)\rangle^{(1)}=\sum_i F_i\int_{-\infty}^{0}dt_1\int_{-\infty}^{\infty}dt_1'\frac{\partial\Gamma(t_1-t_1')}{\partial t_1'}\times\langle\tfrac{1}{2}[Q_i^{(0)}(t_1'),Q_j^{(0)}(t)]_+\rangle^{(0)} \quad (87)$$

or, performing the t_1-integration,

$$\langle Q_j(t)\rangle^{(1)}=-\sum_i F_i\int_{-\infty}^{\infty}dt_1'\Gamma(t_1')\times\langle\tfrac{1}{2}[Q_i^{(0)}(t_1'),Q_j^{(0)}(t)]_+\rangle^{(0)}. \quad (88)$$

In the classical limit $\beta\to 0$, it reduces to Eq. (37) (with $\langle q_i\rangle^{(0)}=0$).

Throughout this and the preceding two sections we have been explicitly concerned with the driven response of standard thermodynamic variables such as volume, number of particles, magnetic moment, etc. However,

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

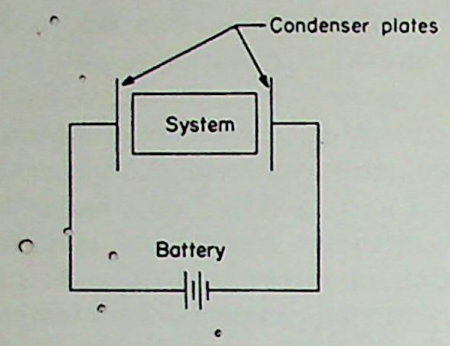


FIG. 2.

in the important case of transport processes one is interested in driven generalized currents rather than in conventional thermodynamic parameters. The special considerations required to treat currents, and particularly to treat the steady state, have been carefully discussed elsewhere.[11] Before concluding the discussion of the linear theory of irreversibility, we indicate briefly the manner of formulation of the theory in terms of driven currents.

We consider the case of electrical conduction, for example, by assuming the specific Hamiltonian

$$H(t)=H^{(0)}+\sum_{i=1}^{3}\mathcal{E}_i(t)Q_i=H^{(0)}+e\sum_{i=1}^{3}\mathcal{E}_i(t)\sum_{\mu}x_{i\mu} \quad (89)$$

where $\mathcal{E}_i(t)$ is the applied electric field in the ith direction, and $x_{i\mu}$ is the ith displacement component of μth charged particle. The component J_i of the current operator $\mathbf{J}$ is simply the time derivative of the operator Q_i.

$$J_i=\dot{Q}_i=e\sum_{\mu}\dot{x}_{i\mu}. \quad (90)$$

In our interpretation heretofore we would have visualized the Hamiltonian (89) as applying to the physical situation illustrated in Fig. 2. The field is applied to the sample by condenser plates which are *not* in physical contact with the sample. The state asymptotically approached after imposition of a step-function force is one with zero current. An alternative interpretation arises if we formally impose periodic boundary conditions on the particle wave functions in the system. A step-function applied force then leads asymptotically to a steady-state current. The formalism is essentially unchanged, but the trace of any operator implies a summation over an entirely different Hilbert space than has been implied heretofore.

The first-order current response $\langle J_j(t)\rangle^{(1)}$ is given, from Eq. (15), by

$$\langle J_j(t)\rangle^{(1)}=-\frac{1}{i\hbar}\sum_{i=1}^{3}\int_{-\infty}^{t}dt_1\mathcal{E}_i(t_1)\langle[Q_i^{(0)}(t_1),J_j^{(0)}(t)]_-\rangle^{(0)}$$

$$=-\frac{1}{i\hbar}\sum_{i=1}^{3}\int_{-\infty}^{t}dt_1\mathcal{E}_i(t_1)\,{}^{J}\phi_{ij}^{(1)}(t-t_1) \quad (91)$$

where the current after-effect function ${}^{J}\phi_{ij}^{(1)}(t)$ is

$${}^{J}\phi_{ij}^{(1)}(t)=-\frac{1}{i\hbar}\langle[Q_i,J_j^{(0)}(t)]_-\rangle^{(0)}. \quad (92)$$

The fluctuation-dissipation theorem relating the first-order current response to the second moment $\langle\frac{1}{2}[J_i,J_j^{(0)}(t)]_+\rangle^{(0)}$ of the spontaneous equilibrium current fluctuations is readily obtained using the techniques employed previously in connection with the extensive parameter displacements Q_i. The results, in the spectral and temporal representations, respectively, are

$${}^{(s)}\langle\tfrac{1}{2}[J_i,J_j^{(0)}(t)]_+\rangle^{(0)}$$

$$=-\frac{2}{\pi}\int_0^{\infty}d\omega\cos\omega tE^{(1)}(\omega;\beta)\,\mathrm{Re}\,{}^{(s)}Y_{ij}(\omega), \quad (93)$$

$${}^{(a)}\langle\tfrac{1}{2}[J_i,J_j^{(0)}(t)]_+\rangle^{(0)}$$

$$=\frac{2}{\pi}\int_0^{\infty}d\omega\sin\omega tE^{(1)}(\omega;\beta)\,\mathrm{Im}\,{}^{(a)}Y_{ij}(\omega), \quad (94)$$

$${}^{J}\phi_{ij}^{(1)}(t)=\int_{-\infty}^{\infty}dt'\Gamma(t-t')\langle\tfrac{1}{2}[J_i,J_j^{(0)}(t')]_+\rangle^{(0)}. \quad (95)$$

Equations (93) and (94) follow immediately from Eqs. (71) and (72), if we replace the operators Q_i by their time derivatives J_i, which simply removes the factor $(1/\omega)^2$ in the integrand. Similarly, Eq. (95) corresponds to Eq. (82) for $\phi_{ij}^{(1)}(t)$.

Finally, the above analysis of driven currents can be justified by another consideration, which is perhaps more physical than the artifice of periodic boundary conditions applied to the system in Fig. 2. We consider a time-dependent magnetic field $\mathcal{H}(t)$ imposed axially through a toroidal conductor, as shown in Fig. 3. The induced current in the toroid will be driven by a tangential electric field $\mathcal{E}(t)=-\dot{\mathbf{A}}(t)$, $\mathbf{A}(t)$ being the vector potential associated with $\mathcal{H}(t)$. The Hamiltonian appropriate to this situation is

$$H(t)=H^{(0)}+A(t)J \quad (96)$$

where J is the operator corresponding to the electrical current around the toroid.

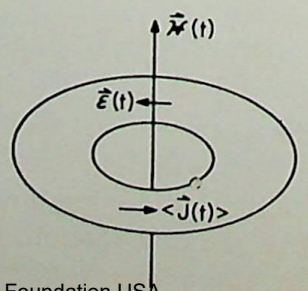


FIG. 3.

[11] See W. Kohn and J. M. Luttinger, Phys. Rev. 108, 590 (1957).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

The first-order response $\langle J(t)\rangle^{(1)}$ to the perturbation $A(t)J$ is then

$$\langle J(t)\rangle^{(1)} = -\frac{1}{i\hbar}\int_{-\infty}^{t} dt_1 A(t_1)\langle[J^{(0)}(t_1), J^{(0)}(t)]_-\rangle^{(0)}. \quad (97)$$

$\langle J(t)\rangle^{(1)}$ can be rewritten in terms of the induced electric field $\mathcal{E}(t)$ by integrating Eq. (97) by parts.

$$\langle J(t)\rangle^{(1)} = -\frac{1}{i\hbar}A(t)\int_{-\infty}^{t} dt_1\langle[J^{(0)}(t_1), J^{(0)}(t)]_-\rangle^{(0)}$$

$$-\frac{1}{i\hbar}\int_{-\infty}^{t} dt_1 \mathcal{E}(t_1)\int_{-\infty}^{t_1} dt_1'\langle[J^{(0)}(t_1'), J^{(0)}(t)]_-\rangle^{(0)} \quad (98)$$

where we have let $\dot{A}(t_1) = -\mathcal{E}(t_1)$.

The term involving $\mathcal{E}(t)$ is the physically interesting one, the integrated term corresponding simply to the accumulation of magnetic field required to sustain the driving electric field. By ignoring the latter term, the physical situation is precisely that which would obtain if the process were driven by a battery placed in the circuit rather than by the magnetic field $\mathcal{H}(t)$. For this case, $\langle J(t)\rangle^{(1)}$ reduces to

$$\langle J(t)\rangle^{(1)} = -\frac{1}{i\hbar}\int_{-\infty}^{t} dt_1 \mathcal{E}(t_1)$$

$$\times\int_{-\infty}^{t_1} dt_1'\langle[J^{(0)}(t_1'), J^{(0)}(t)]_-\rangle^{(0)}. \quad (99)$$

By Eq. (90) we identify

$$Q_i^{(0)}(t_1) = \int_{-\infty}^{t_1} dt_1' J^{(0)}(t_1'). \quad (100)$$

Therefore,

$$\langle J(t)\rangle^{(1)} = -\frac{1}{i\hbar}\int_{-\infty}^{t} dt_1 \mathcal{E}(t_1)\langle[Q^{(0)}(t_1), J^{(0)}(t)]_-\rangle^{(0)} \quad (101)$$

which is identical with Eq. (91) for the case of one-dimensional electrical conduction.

8. APPLICATIONS OF THE FIRST-ORDER THEORY

Several applications of the foregoing first-order theory are now mentioned briefly.

In their original paper on the fluctuation-dissipation theorem, Callen and Welton[4] discussed the relation of that theorem to the energy density in an isotropic radiation field. The impedance of a charged particle driven by a periodic electric field exhibits a dissipative term arising from the radiation damping force. According to the fluctuation-dissipation theorem (78), this implies the existence of a random fluctuating electric field exerted by the vacuum on the free particle. The energy density of this fluctuating field is found to be just the familiar Planck radiation density.

Van Vliet[12] has recently employed the fluctuation-dissipation theorem to discuss the equilibrium charge carrier fluctuations in semiconducting materials. A simple admittance matrix corresponding to a linear RC network is introduced, the resistances being expressed in terms of transition rates between different groups of carrier levels. The fluctuation-dissipation theorem thereby yields the second moments of the equilibrium carrier fluctuations in terms of the thermal generation-recombination process. The charge carrier fluctuations in turn give rise to a contribution to the driven noise, to which further reference is made in Sec. 12.

In addition to the fluctuation-dissipation theorem and the spectral reciprocity, Kubo[9] points out that general proofs of certain sum rules can be obtained from irreversible thermodynamic considerations. Thus, for the case of electrical conductivity in a system of interacting particles in an applied magnetic field, he finds that the frequency integrals of $\mathrm{Re}^{(s)}Y_{ij}(\omega)$ and $\omega\,\mathrm{Im}^{(a)}Y_{ij}(\omega)$ are given by

$$\frac{2}{\pi}\int_0^{\infty} d\omega\,\mathrm{Re}^{(s)}Y_{ij}(\omega) = \sum_r \frac{e^2 n_r}{m_r}\delta_{ij}, \quad (102)$$

$$\frac{2}{\pi}\int_0^{\infty} d\omega\,\omega\,\mathrm{Im}^{(a)}Y_{ij}(\omega) = \sum_r \frac{e_r^3 n_r}{m_r^2 c}\mathcal{H}_z. \quad (103)$$

n_r, m_r, and e_r are the number, mass, and charge, respectively, of the rth type of particle, and $\mathcal{H}_z$ is the z-directed applied magnetic field. Analogous sum rules can be derived for the magnetic susceptibility matrix.

H. Mori[13] has applied the fluctuation-disspation theorem to the analysis of transport processes in fluids. The coupling between the slow macroscopic relaxation of the system and the rapid microscopic fluctuations is shown to be responsible for the dissipation. Thus, the coefficients of viscosity, thermal conductivity, and diffusion can be computed in terms of the equilibrium fluctuations of the thermodynamic fluxes.

9. THE SECOND-ORDER RESPONSE IN A GENERAL PROCESS

In Secs. 5 through 8 the first-order theory of irreversible thermodynamics was reviewed, showing the relationship of $\langle Q_i(t)\rangle^{(1)}$ to the equilibrium fluctuations. Sections 9 through 14 are devoted to an extension of the fluctuation-dissipation concept to the driven second moment $\langle\frac{1}{2}[Q_i(t), Q_j(t+\tau)]_+\rangle$ and to the second- and higher-order terms in the driven response $\langle Q_i(t)\rangle$. A number of interrelationships among these quantities and the equilibrium fluctuations are established.

[12] K. M. Van Vliet, Phys. Rev. **110**, 50 (1958).
[13] H. Mori, Phys. Rev. **112**, 1829 (1958).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

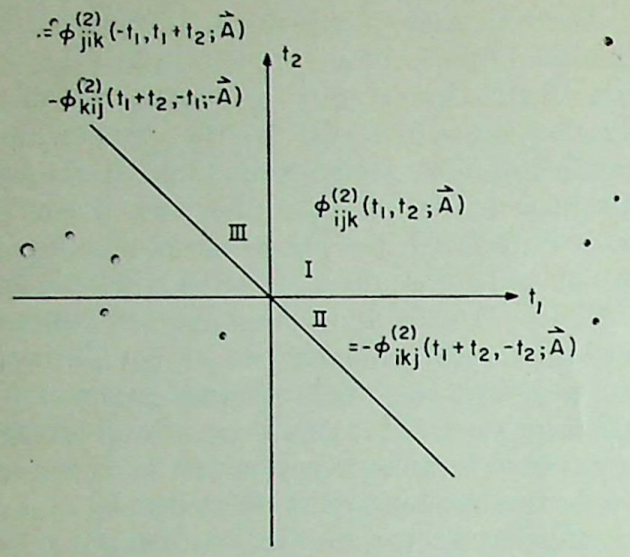


FIG. 4.

Consider the second-order response $\langle Q_i(t)\rangle^{(2)}$. From Eq. (16),

$$\langle Q_i(t)\rangle^{(2)}=\frac{1}{2}\left(-\frac{1}{i\hbar}\right)^2\sum_{jk}\Big\{\int_{-\infty}^{t}dt_1F_j(t_1)\int_{-\infty}^{t_1}dt_2F_k(t_2)$$
$$\times\langle[[Q_k^{(0)}(t_2),[Q_j^{(0)}(t_1),Q_i^{(0)}(t)]_-]_-]_-\rangle^{(0)}$$
$$+\int_{-\infty}^{t}dt_1F_j(t_1)\int_{t_1}^{t}dt_2F_k(tz)\langle[[Q_j^{(0)}(t_1),$$
$$\times[Q_k^{(0)}(t_2),Q_i^{(0)}(t)]_-]_-]_-\rangle^{(0)}\Big\} \quad (104)$$

where we have inverted the order of integration in the second term. The second-order after-effect function $\phi_{ijk}^{(2)}(t_1,t_2)$ is defined by rewriting Eq. (104) as

$$\langle Q_i(t)\rangle^{(2)}=\frac{1}{2}\sum_{jk}\Big\{\int_{-\infty}^{t}dt_1F_j(t_1)\int_{-\infty}^{t_1}dt_2F_k(t_2)$$
$$\times\phi_{kji}^{(2)}(t_1-t_2,\,t-t_1)$$
$$+\int_{-\infty}^{t}dt_1F_j(t_1)\int_{t_1}^{t}dt_2F_k(t_2)$$
$$\times\phi_{jki}^{(2)}(t_2-t_1,\,t-t_2)\Big\} \quad (105)$$

whence

$$\phi_{ijk}^{(2)}(t_1,t_2)$$
$$=\left(-\frac{1}{i\hbar}\right)^2\langle[Q_i,[Q_j^{(0)}(t_1),\,Q_k^{(0)}(t_1+t_2)]_-]_-\rangle^{(0)}. \quad (106)$$

$\phi_{ijk}^{(2)}(\tau;\,t-\tau)$ is the response $\langle Q_k(t)\rangle^{(2)}$ at time $t(>0)$ to a δ-function force F_i applied at time zero, and a δ-function force F_j applied at time $\tau(>0)$. Consequently, the first term in Eq. (105) characterizes the contribution to $\langle Q_i(t)\rangle^{(2)}$ arising from the application of $F_k(t_2)$ prior to $F_j(t_1)$, while the second term characterizes the contribution arising from the application of $F_k(t_2)$ subsequent to $F_j(t_1)$.

Although Eq. (105) defines $\phi_{ijk}^{(2)}(t_1,t_2)$ only for positive t_1 and t_2, we accept Eq. (106) as the formal definition of $\phi_{ijk}^{(2)}(t_1,t_2)$ for arbitrary values of t_1 and t_2. The symmetry properties of $\phi_{ijk}^{(2)}(t_1,t_2)$ permit us to evaluate this function for arbitrary times in terms of its measured values for positive times.

We first observe that $\phi_{ijk}^{(2)}(t_1,t_2)$ is invariant with respect to simultaneous reversal of all times and the applied magnetic vector potential $\mathbf{A}$, the argument being identical to that given below Eq. (45),

$$\phi_{ijk}^{(2)}(-t_1,\,-t_2;\,-\mathbf{A})=\phi_{ijk}^{(2)}(t_1,t_2;\,\mathbf{A}). \quad (107)$$

Further, $\phi_{ijk}^{(2)}(t_1,t_2)$ can be written in either of the following forms.

$$\phi_{ijk}^{(2)}(t_1,t_2)=\left(-\frac{1}{i\hbar}\right)^2\{-\langle[Q_j^{(0)}(t_1),$$
$$\times[Q_k^{(0)}(t_1+t_2),\,Q_i]_\mp]_\mp\rangle^{(0)}$$
$$+\langle[[Q_i,\,Q_j^{(0)}(t_1)]_\mp,$$
$$\times Q_k^{(0)}(t_1+t_2)]_\mp\rangle^{(0)}\}. \quad (108)$$

These forms follow by writing out all terms in the double commutator of Eq. (106) and appropriately regrouping. Consider the double commutator form of Eq. (108), which states that

$$\phi_{ijk}^{(2)}(t_1,t_2)+\phi_{jki}^{(2)}(t_2,\,-t_1-t_2)$$
$$+\phi_{kij}^{(2)}(-t_1-t_2,\,t_1)=0. \quad (109)$$

This cyclic relationship corresponds to Eq. (43) for $\phi_{ij}^{(1)}(t)$.

We now return to our observation that the symmetries (107) and (109) permit us to evaluate $\phi_{ijk}^{(2)}(t_1,t_2)$ for arbitrary times from its measured value for positive times. Consider the t_1t_2 plane shown in Fig. 4, which we have divided into sectors. The value of $\phi_{ijk}^{(2)}(t_1,t_2;\,\mathbf{A})$ in sector I is obtained by direct measurement.

The value in sector II (for which $t_1\gtrless -t_2>0$) is obtained by noting that

$$\phi_{ijk}^{(2)}(t_1,t_2;\,\mathbf{A})=-\phi_{ijk}^{(2)}(t_1+t_2;\,-t_2;\,\mathbf{A}).$$

Sector III is determined by rewriting Eq. (109), interchanging the indice ik in the second term and reversing the vector potential $\mathbf{A}$ in the third term.

$$\phi_{ijk}^{(2)}(t_1,t_2;\,\mathbf{A})-\phi_{jik}^{(2)}(-t_1,\,t_1+t_2;\,\mathbf{A})$$
$$+\phi_{kij}^{(2)}(t_1+t_2,\,-t_1;\,-\mathbf{A})=0. \quad (110)$$

Then for $t_2>-t_1>0$ the second and third terms are measurable, thereby determining the value of the first term in sector IV.

Equation (107) reflects the known values into the remaining half-plane.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

We now discuss the relationship which exists between $\phi_{ijk}^{(2)}(t_1,t_2)$ and the equilibrium fluctuations. Consider the form of Eq. (108), for $\phi_{ijk}^{(2)}(t_1,t_2)$, involving the double anticommutators. This can be rewritten

$$\phi_{ijk}^{(2)}(t_1,t_2)=4\left(-\frac{1}{i\hbar}\right)^2[\Psi_{kij}^{(0)}(-t_1-t_2,t_1) - \Psi_{jki}^{(0)}(t_2,-t_1-t_2)] \quad (111)$$

where

$$\Psi_{ijk}^{(0)}(t_1,t_2)=\tfrac{1}{4}\langle[Q_i,[Q_j^{(0)}(t_1),Q_k^{(0)}(t_1+t_2)]_+]_+\rangle^{(0)}. \quad (112)$$

Since $\Psi_{ijk}^{(0)}(t_1,t_2)$ is the equilibrium expectation value of a symmetrized product of the operators Q_i, $Q_j^{(0)}(t_1)$, and $Q_k^{(0)}(t_1+t_2)$, it is a third correlation moment among the equilibrium fluctuations of the variables corresponding to these operators.

However, referring back to Eq. (108) to identify the two third equilibrium correlation moments in Eq. (111), we see that each involve precisely the same operators at precisely the same times, although the order of symmetrization is different! The two distinct third moments correspond in principle to different ways of measuring the correlation, as can be seen from the following general considerations.

Since there is only one way of symmetrizing a product of two non-commuting operators, it is possible to write a unique quantum-mechanical expression for the second equilibrium correlation moment [see Eq. (17)]. However, quantum mechanics furnishes no such unique *a priori* prescription for symmetrizing a product of three (or more) operators.[14] Thus, for example, in Eq. (112) we introduced the equilibrium symmetrized quantity $\Psi_{ijk}^{(0)}(t_1,t_2)$ containing four permutations of the operator product $Q_iQ_j^{(0)}(t_1)Q_k^{(0)}(t_1,t_2)$, while we can also construct the fully symmetrized form $\bar{\Psi}_{ijk}^{(0)}(t_1,t_2)$ containing all six permutations.

$$\begin{aligned}\bar{\Psi}_{ijk}^{(0)}(t_1,t_2) &= \tfrac{1}{6}\langle Q_iQ_j^{(0)}(t_1)Q_k^{(0)}(t_1+t_2) \\ &\quad + Q_j^{(0)}(t_1)Q_k^{(0)}(t_1+t_2)Q_i \\ &\quad + Q_k^{(0)}(t_1+t_2)Q_iQ_j^{(0)}(t_1)\rangle^{(0)} \\ &\quad + (\text{complex conjugate}). \end{aligned} \quad (113)$$

Each possible symmetrized arrangement corresponds to some particular experimental measurement. Consider an experiment in which three detectors monitor the variables Q_i, Q_j, and Q_k, and feed their signals into a counter. Appropriate time delay circuits are inserted between the Q_i and Q_j detectors and the counter such that the counter makes a single measurement of the desired product. Since all three signals are handled in a completely symmetrical fashion, this experiment measures the fully symmetrized operator product $\bar{\Psi}_{ijk}^{(0)}(t_1,t_2)$.

Alternately, suppose the Q_k signal and the delayed Q_j signal are fed into a multiplier, and the multiplied signal is fed into the counter along with the delayed Q_i signal. The counter makes a simultaneous measurement of the product of its two inputs. In this case Q_j and Q_k are treated symmetrically, as are their product and Q_i, and the experiment measures the quantity $\Psi_{ijk}^{(0)}(t_1,t_2)$ of Eq. (112).

In the classical limit (zero order in β) the two correlation moments become identical, as discussed later in Sec. 13.

Returning to Eq. (111), we see that this expression constitutes a thermodynamic relationship between the second-order response, characterized by the second-order aftereffect function $\phi_{ijk}^{(2)}(t_1,t_2)$, and the equilibrium fluctuations, characterized by the difference between the two operationally distinct third correlation moments $\Psi_{kij}^{(0)}(-t_1-t_2,t_1)$ and $\Psi_{jki}^{(0)}(t_2,-t_1-t_2)$.

10. FIRST-ORDER NOISE IN A GENERAL PROCESS

In this section we consider the first-order term $\langle[Q_i(t),Q_j(t+\tau)]_+\rangle^{(1)}$ in the driven second moment, establishing its relationship to the equilibrium fluctuations. The relationship between the first-order driven noise and the second-order response is also discussed.

The relationship of the first-order driven second moment to the equilibrium fluctuations is conveniently approached using Eq. (20) for $\langle\frac{1}{2}[Q_i(t),Q_j(t+\tau)]_+\rangle^{(1)}$. The quantity

$$-(1/2i\hbar)\langle[Q_k^{(0)}(t_1),[Q_i^{(0)}(t),Q_j^{(0)}(t+\tau)]_+]_-\rangle^{(0)}$$

appearing in the first term of Eq. (20) is the noise response $\langle\frac{1}{2}[Q_i(t),Q_j(t+\tau)]_+\rangle^{(1)}$ to an applied force $F_k(t)=\delta(t-t_1)$, $t_1<t<t+\tau$. As such, it characterizes the noise contribution arising from the application of $F_k(t_1)$ prior to time t. This noise response function can be readily symmetrized and is the quantity of primary physical interest.

On the other hand, the quantity

$$-(1/2i\hbar)\langle[Q_i^{(0)}(t),[Q_k^{(0)}(t_1),Q_j^{(0)}(t+\tau)]_-]_+\rangle^{(0)}$$

appearing in the second term of Eq. (20) is the noise response to the force $F_k(t)=\delta(t-t_1)$, $t<t_1<t+\tau$. As such it characterizes the noise contribution arising from the application of $F_k(t_1)$ in the time interval t to $t+\tau$, during which the noise is being measured. We denote this function by $\theta_{ikj}^{(1)}(t_1-t,t+\tau-t_1)$. Thus,

$$\theta_{ijk}^{(1)}(t_1,t_2) = -\frac{1}{2i\hbar}\langle[Q_i,[Q_j^{(0)}(t_1),Q_k^{(0)}(t_1+t_2)]_-]_+\rangle^{(0)}. \quad (114)$$

As we shall see, $\theta_{ijk}^{(1)}(t_1,t_2)$ is not of particular physical interest; we discuss it briefly at the end of this section in connection with the second-order response.

The first equilibrium expectation value appearing on the right-hand side of Eq. (20) can be symmetrized by

[14] J. R. Shewell, Am. J. Phys. 27, 16 (1959).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

invoking the basic relationship (82) between commutators and anticommutators. Replacing t by $(t-t_1)$ and introducing the transformation $t_1'=t-t'$ gives Eq. (82) in the form

$$-\frac{1}{i\hbar}\langle[Q_k^{(0)}(t_1),Q_i^{(0)}(t)]_-\rangle^{(0)}$$

$$=-\int_{-\infty}^{\infty}dt_1'\Gamma(t_1-t_1')\langle\tfrac{1}{2}[\dot{Q}_k^{(0)}(t_1'),Q_i^{(0)}(t)]_+\rangle^{(0)}. \quad (115)$$

If we replace the operator Q_i by the anticommutator $\frac{1}{2}[Q_i,Q_j^{(0)}(\tau)]_+$, this becomes

$$-\frac{1}{2i\hbar}\langle[Q_k^{(0)}(t_1),[Q_i^{(0)}(t),Q_j^{(0)}(t+\tau)]_+]_-\rangle^{(0)}$$

$$=-\int_{-\infty}^{\infty}dt_1'\Gamma(t_1-t_1')\langle\tfrac{1}{4}[\dot{Q}_k^{(0)}(t_1'),$$
$$\times[Q_i^{(0)}(t),Q_j^{(0)}(t+\tau)]_+]_+\rangle^{(0)}. \quad (116)$$

The result (116) can be substituted directly into Eq. (20), along with the definition (114) of $\theta_{ijt}^{(1)}(t_1,t_2)$, to yield the first-order driven noise in the form

$$\langle\tfrac{1}{2}[Q_i(t),Q_j(t+\tau)]_+\rangle^{(1)}$$

$$=-\sum_k\int_{-\infty}^{t}dt_1F_k(t_1)\int_{-\infty}^{\infty}dt_1'\Gamma(t_1-t_1')$$
$$\times\langle\tfrac{1}{4}[\dot{Q}_k^{(0)}(t_1'),[Q_i^{(0)}(t),Q_j^{(0)}(t+\tau)]_+]_+\rangle^{(0)}$$
$$+\sum_k\int_t^{t+\tau}dt_1F_k(t_1)\theta_{ikj}^{(1)}(t_1-t,t+\tau-t_1). \quad (117)$$

As discussed in the preceding section, the symmetrized equilibrium expectation value

$$\langle\tfrac{1}{4}[\dot{Q}_k^{(0)}(t_1'),[Q_i^{(0)}(t),Q_j^{(0)}(t+\tau)]_+]_+\rangle^{(0)}$$

is one macroscopically observable form of the third correlation moment among the spontaneous equilibrium fluctuations of the variables corresponding to the operators $\dot{Q}_k^{(0)}(t_1')$, $Q_i^{(0)}(t)$, and $Q_j^{(0)}(t+\tau)$.

A thermodynamic relationship between the intensity of the first-order driven noise and the equilibrium fluctuations follows immediately from Eq. (117). Letting $\tau=0$, the second term in Eq. (117) vanishes, leaving

$$\langle\tfrac{1}{2}[Q_i(t),Q_j(t)]_+\rangle^{(1)}$$

$$=-\sum_k\int_{-\infty}^{t}dt_1F_k(t_1)\int_{-\infty}^{\infty}dt_1'\Gamma(t_1-t_1')$$
$$\times\langle\tfrac{1}{4}[\dot{Q}_k^{(0)}(t_1'),[Q_i^{(0)}(t),Q_j^{(0)}(t)]_+]_+\rangle^{(0)}. \quad (118)$$

Equation (118) expresses the first-order driven noise intensity $\langle\frac{1}{2}[Q_i(t),Q_j(t)]_+\rangle^{(1)}$ in terms of the equilibrium third fluctuation moment

$$\langle\tfrac{1}{4}[\dot{Q}_k^{(0)}(t_1'),[Q_i^{(0)}(t),Q_j^{(0)}(t)]_+]_+\rangle^{(0)}.$$

Returning to the more general quantity

$$\langle\tfrac{1}{2}[Q_i(t),Q_j(t+\tau)]_+\rangle^{(1)},$$

for most purposes one is interested in measuring this driven noise only for processes in which the imposed forces are slowly varying over the time interval during which the noise is being measured. We therefore assume that $F_k(t_1)\simeq$constant$\equiv F_k$ in the interval t to $(t+\tau)$ although, of course, $F_k(t_1)$ is arbitrary for $t_1<t$. Further, we decompose $F_k(t_1)$ into two components according to

$$F_k(t_1)=F_k+\Delta F_k(t_1). \quad (119)$$

The contribution to $\langle\frac{1}{2}[Q_i(t),Q_j(t+\tau)]_+\rangle^{(1)}$ arising from the time-dependent force component $\Delta F_k(t_1)$ is

$$-\sum_k\int_{-\infty}^{t}dt_1\Delta F_k(t_1)\int_{-\infty}^{\infty}dt_1'\Gamma(t_1-t_1')$$
$$\times\langle\tfrac{1}{4}[\dot{Q}_k^{(0)}(t_1'),[Q_i^{(0)}(t),Q_j^{(0)}(t+\tau)]_+]_+\rangle^{(0)} \quad (120)$$

since the integral $\int_t^{t+\tau}dt_1\Delta F_k(t_1)$ vanishes. It is therefore directly expressible in terms of the equilibrium fluctuations.

The contribution arising from the constant force component F_k, on the other hand, is just the first-order term in the perturbation expansion of the driven second moment with respect to a constant applied force. However, the application of a constant force F_k implies simply a change in the corresponding (equilibrium) intensive parameter associated with the system. Hence, this contribution can also be regarded as a macroscopically observable characteristic of the equilibrium system. We denote it by

$$\sum_k F_k\frac{\partial}{\partial F_k}\langle\tfrac{1}{2}[Q_i,Q_j^{(0)}(\tau)]_+\rangle^{(0)}$$

where the derivative is evaluated at $F_k=0$.

Inserting the quantities (120) and (121) into Eq. (117), we obtain the result

$$\langle\tfrac{1}{2}[Q_i(t),Q_j(t+\tau)]_+\rangle^{(1)}$$

$$=\sum_k\Big\{-\int_{-\infty}^{t}dt_1\Delta F_k(t_1)\int_{-\infty}^{\infty}dt_1'\Gamma(t_1-t_1')$$
$$\times\langle\tfrac{1}{4}[\dot{Q}_k^{(0)}(t_1'),[Q_i^{(0)}(t),Q_j^{(0)}(t+\tau)]_+]_+\rangle^{(0)}$$
$$+F_k\frac{\partial}{\partial F_k}\langle\tfrac{1}{2}[Q_i,Q_j^{(0)}(\tau)]_+\rangle^{(0)}\Big\}. \quad (122)$$

Equation (122) constitutes a thermodynamic relationship between $\langle\frac{1}{2}[Q_i(t),Q_j(t+\tau)]_+\rangle^{(1)}$ and the indicated

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

macroscopically observable characteristics of the equilibrium fluctuations.

We further note that $(\partial/\partial F_k)\langle\frac{1}{2}[Q_i,Q_j^{(0)}(\tau)]_+\rangle^{(0)}$ can be re-expressed in terms of the nonlinear behavior of the system by invoking the fluctuation-dissipation theorem, Eqs. (71) and (72). In order to keep the notation simple, we consider explicitly the case of no applied magnetic field, for which $\langle\frac{1}{2}[Q_i,Q_j^{(0)}(\tau)]_+\rangle^{(0)}$ reduces to

$$\langle\tfrac{1}{2}[Q_i,Q_j^{(0)}(\tau)]_+\rangle^{(0)} = -\frac{2}{\pi}\int_0^\infty d\omega\,\cos\omega\tau E^{(1)}(\omega;\beta)\frac{\mathrm{Re}\,Y_{ij}(\omega)}{\omega^2}. \quad (123)$$

Just as $\langle\frac{1}{2}[Q_i,Q_j^{(0)}(\tau)]_+\rangle^{(0)}$ is a function of the applied forces F_k, $Y_{ij}(\omega)$ also in general depends upon F_k. Consequently, Eq. (122) can be written as

$$\langle\tfrac{1}{2}[Q_i(t),\,Q_j(t+\tau)]_+\rangle^{(0)} = \sum_k\Big\{-\int_{-\infty}^t dt_1\Delta F_k(t_1)\int_{-\infty}^\infty dt_1'\Gamma(t_1-t_1') \times\langle\tfrac{1}{4}[\dot{Q}_k^{(0)}(t_1'),\,[Q_i^{(0)}(t),\,Q_j^{(0)}(t+\tau)]_+]_+\rangle^{(0)} - \frac{2}{\pi}\int_0^\infty d\omega\,\cos\omega\tau E^{(1)}(\omega;\beta)\frac{1}{\omega^2}\mathrm{Re}\frac{\partial Y_{ij}(\omega)}{\partial F_k}\Big\}. \quad (124)$$

The physical significance of the derivative $\partial Y_{ij}(\omega)/\partial F_k$ can be regarded as arising from the nonlinearity of the system in the following way. Most physical systems are nonlinear. Nevertheless, for sufficiently small deviations from a given "operating point" (corresponding to a constant applied force F_k) the linear approximation is adequate. As a consequence of the nonlinearity of the system, however, the admittance matrix must in general be a function of the "operating point." The quantity $\partial Y_{ij}(\omega)/\partial F_k$ specifies the first-order contribution to this dependence on F_k.

Equation (124) is an alternate thermodynamic expression for $\langle\frac{1}{2}[Q_i(t),\,Q_j(t+\tau)]_+\rangle^{(1)}$ to that given in Eq. (122), expressing the first-order driven noise in terms of the equilibrium third moment $\langle\frac{1}{4}[\dot{Q}_k^{(0)}(t_1'), \times[Q_i^{(0)}(t),\,Q_j^{(0)}(t+\tau)]_+]_+\rangle^{(0)}$ and the second-order response function $\partial Y_{ij}(\omega)/\partial F_k$.

In the classical limit, Eqs. (122) and (124) reduce, respectively, to

$$\langle q_i(t)q_j(t+\tau)\rangle^{(1)} = \sum_k\Big\{-\beta\int_{-\infty}^t dt_1\Delta F_k(t_1)\langle\dot{q}_k(t_1)q_i(t)q_j(t+\tau)\rangle^{(0)} + F_k\frac{\partial}{\partial F_k}\langle q_iq_j(\tau)\rangle^{(0)}\Big\}, \quad (125)$$

$$\langle q_i(t)q_j(t+\tau)\rangle^{(1)} = \sum_k\Big\{-\beta\int_{-\infty}^t dt_1\Delta F_k(t_1)\langle\dot{q}_k(t_1)q_i(t)q_j(t+\tau)\rangle^{(0)} - \frac{2}{\pi\beta}F_k\int_0^\infty d\omega\,\cos\omega\tau\frac{1}{\omega^2}\mathrm{Re}\frac{\partial Y_{ij}(\omega)}{\partial F_k}\Big\}. \quad (126)$$

We return finally to a brief discussion of the relationship of the noise response function $\theta_{ijk}^{(1)}(t_1,t_2)$, defined in Eq. (114), to the second-order response. This relationship also stems from the basic relationship between equilibrium commutators and anticommutators given in Eq. (82). Replacing the operator Q_j in Eq. (82) by the commutator $-(1/i\hbar)[Q_j,Q_k^{(0)}(t_2)]_-$ and noting Eqs. (106) and (114) for $\phi_{ijk}^{(2)}(t_1,t_2)$ and $\theta_{ijk}^{(1)}(t_1,t_2)$, respectively, we find that

$$\phi_{ijk}^{(2)}(t_1,t_2) = \int_{-\infty}^\infty dt_1'\Gamma(t_1-t_1')\frac{\partial}{\partial t_1'}\theta_{ijk}^{(1)}(t_1',t_2). \quad (127)$$

The spectral form of Eq. (127) is found to be

$$\theta_{ijk}^{(1)}(t_1,t_2) = \frac{1}{2\pi}\int_{-\infty}^\infty d\omega_1\int_{-\infty}^\infty d\omega_2 e^{i\omega_1t_1}e^{i\omega_2t_2} \times E^{(1)}(\omega;\beta)\frac{L_{ijk}^{(2)}(\omega_1,\omega_2)}{i\omega_1} \quad (128)$$

where $L_{ijk}^{(2)}(\omega_1,\omega_2)$ is the double Fourier transforms of $\phi_{ijk}^{(2)}(t_1,t_2)$ defined by

$$\phi_{ijk}^{(2)}(t_1,t_2) = \frac{1}{2\pi}\int_{-\infty}^\infty d\omega_1\int_{-\infty}^\infty d\omega_2 e^{i\omega_1t_1}e^{i\omega_2t_2} \times L_{ijk}^{(2)}(\omega_1,\omega_2). \quad (129)$$

Equations (128) and (129) have the following formal implications. Using Eq. (114) for $\theta_{ijk}^{(1)}(t_1,t_2)$, Eq. (19) for $\langle\frac{1}{2}[Q_i(t),\,Q_j(t+\tau)]_+\rangle^{(1)}$ can be written in the form

$$\langle\tfrac{1}{2}[Q_i(t),\,Q_j(t+\tau)]_+\rangle^{(1)} = \sum_k\Big\{\int_{-\infty}^t dt_1F_k(t_1)\theta_{jki}^{(1)}(t_1-t-\tau,\,t-t_1) + \int_{-\infty}^{t+\tau} dt_1F_k(t_1)\theta_{ikj}^{(1)}(t_1-t,\,t+\tau-t_1)\Big\}. \quad (130)$$

Since Eqs. (127) and (128) express the function $\theta_{ijk}^{(1)}(t_1,t_2)$ in terms of the macroscopically observable second-order aftereffect function $\phi_{ijk}^{(2)}(t_1,t_2)$, Eq. (130) permits us to compute $\langle\frac{1}{2}[Q_i(t),\,Q_j(t+\tau)]_+\rangle^{(0)}$ from suitable measurements of the second-order response.

11. FLUCTUATION SYMMETRY

In the preceding two sections we have established the basic interrelationship among the second-order response,

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

the first-order second moment, and the equilibrium third moment.

In order better to appreciate this basic interrelationship, it is of interest to examine its consistency from the point of view of symmetry. For this purpose it is convenient to represent the response $\langle Q_i(t)\rangle$ to a set of applied forces $F_j(t)$ by the symbolic expansion

$$\langle Q_i(t)\rangle = \chi_i^{(0)} + \sum_j \mathfrak{F}_j \chi_{ji}^{(0)} + \tfrac{1}{2}\sum_{jk} \mathfrak{F}_j \mathfrak{F}_k \chi_{jki}^{(2)} + \cdots . \quad (131)$$

The response functions χ are suitable combinations of the aftereffect functions defined previously, while the $\mathfrak{F}_j$ are integral operators, linear functionals of the forces $F_j(t)$ acting on the χ's. $\chi_i^{(0)}$ denotes the equilibrium expectation value $\langle Q_i\rangle^{(0)}$. Similarly, the driven second moment $\langle \tfrac{1}{2}[Q_i(t), Q_j(t+\tau)]_+\rangle$ can be represented by the expansion

$$\langle \tfrac{1}{2}[Q_i(t), Q_j(t+\tau)]_+\rangle = \xi_{ij}^{(0)} + \sum_k \mathfrak{F}_k' \xi_{kij}^{(1)} + \tfrac{1}{2}\sum_{kl} \mathfrak{F}_k' \mathfrak{F}_l' \xi_{klij} + \cdots \quad (132)$$

where the ξ are suitable noise response functions, and the $\mathfrak{F}_k'$ are appropriate integral operators.

The physical symmetry of many systems is such that reversal of all forces simply reverses all responses $\langle Q_i(t)\rangle - \langle Q_i\rangle^{(0)}$. For such systems only the odd terms in Eq. (131), and only the even terms in Eq. (132) can exist.

The physical symmetry referred to above also has obvious implications for the equilibrium fluctuations. Consider particularly the third equilibrium correlation moment $\langle q_i q_j(t_1) q_k(t_2)\rangle^{(0)}$. This moment is defined in terms of an integral involving the equilibrium joint probability distribution $W_3^{(0)}(q_i; q_j,t_1; q_k,t_2)$. For many systems the physical symmetry implies that this joint probability distribution is unchanged if each of the q's is replaced by its negative. For such a "fluctuation-symmetric" system, all odd equilibrium correlation moments vanish.

The relations which were proven among the second-order response, the first-order noise, and the equilibrium third moments indicate that the physical symmetries referred to in the two preceding paragraphs are equivalent, as we might intuitively expect. A system which is fluctuation-symmetric, with no odd equilibrium moments, exhibits no second-order response, and no first-order noise.

A homogeneous system, symmetric under spatial inversion, is fluctuation-symmetric with respect to its transport properties. In such systems, electron, phonon, or other currents can exhibit no first-order noise.

With fluctuation-symmetric systems, it is necessary to go to second order to obtain a contribution to the driven noise. Such second-order driven noise can be appreciable in magnitude, and, in fact, is easily observed in semiconductors. By analogy with Eqs. (111), (124), and (127), the second-order noise may be presumed to depend generally in some complicated way upon the third-order response and the equilibrium fourth moment. In Sec. 12, we give a limited discussion of one contribution to second-order noise.

On the other hand, there are many systems which do not obey fluctuation-symmetry and which therefore may exhibit first-order driven noise. Rectifiers, for example, because of their pronounced asymmetry with respect to current flow, necessarily possess significant equilibrium third moments.

Finally, it is possible for a system to be fluctuation-symmetric with respect to some of its variables but not with respect to others. Thus, for example, a p-n junction is fluctuation-symmetric with respect to current in the plane of the junction but fluctuation-asymmetric with respect to current perpendicular to this plane. Another example would be a bulk solid, which we have previously mentioned as being fluctuation-symmetric with respect to its transport properties. Such a system would not in general be fluctuation symmetric with respect to its thermodynamic extensive variables such as energy, volume, or the number of particles in the conduction band.

12. SECOND-ORDER DRIVEN NOISE

Although we do not undertake a complete discussion of the second-order noise in this paper, it is of interest to indicate how the general theory would apply in a specific physical situation. As an example, we consider the steady-state thermal generation-recombination noise in semiconductors.

We consider explicitly the second-order driven current noise $\langle \tfrac{1}{2}[J_i(t), J_j(t+\tau)]_+\rangle^{(2)}$. In accordance with the recipe developed in Sec. 3 for computing driven second moments in terms of the equilibrium system, we have that

$$\begin{aligned}\langle \tfrac{1}{2}[J_i(t), J_j(t+\tau)]_+\rangle^{(2)} &= \tfrac{1}{2}\{\langle [J_i^{(0)}(t), J_j^{(2)}(t+\tau)]_+\rangle^{(0)} \\ &\quad + \langle [J_i^{(1)}(t), J_j^{(1)}(t+\tau)]_+\rangle^{(0)} \\ &\quad + \langle [J_i^{(2)}(t), J_j^{(0)}(t+\tau)]_+\rangle^{(0)}\}. \end{aligned} \quad (133)$$

Although the first and third terms also contribute to $\langle \tfrac{1}{2}[J_i(t), J_j(t+\tau)]_+\rangle^{(2)}$, we focus attention on the term $\langle \tfrac{1}{2}[J_i^{(1)}(t), J_j^{(1)}(t+\tau)]_+\rangle^{(0)}$, which is subject to clear physical interpretation.

The first-order driven current operator $J_i^{(1)}(t)$ is obtained in accordance with the discussion at the end of Sec. 7.

$$J_i^{(1)}(t) = -\frac{1}{i\hbar}\sum_k \int_{-\infty}^{t} dt_1 \mathcal{E}_k(t_1)[Q_k^{(0)}(t_1), J_i^{(0)}(t)]_-. \quad (134)$$

In the steady state, the electric field $\mathcal{E}_k(t)$ is constant,

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

and Eq. (134) reduces to

$$J_i^{(1)}(t)=-\frac{1}{ih}\sum_k \mathcal{E}_k \exp\left[i\frac{H^{(0)}t}{\hbar}\right] \times\int_{-\infty}^{0} dt_1[Q_k^{(0)}(t_1),J_i]_- \exp\left[-i\frac{H^{(0)}t}{\hbar}\right] \quad (135)$$

where we have let $t_1'=t-t_1$ and extracted the resulting t dependence of the commutator as indicated.

We define the unperturbed Heisenberg operator $\sigma_{ki}^{(0)}(t)$ corresponding to the kith element of the conductivity matrix by rewriting Eq. (135) as

$$J_i^{(1)}(t)=\sum_k \mathcal{E}_k\sigma_{ki}^{(0)}(t) \quad (136)$$

whence

$$\sigma_{ki}^{(0)}(t)=-\frac{1}{ih}\exp\left[i\frac{H^{(0)}t}{\hbar}\right] \times\int_{-\infty}^{0} dt_1[Q_k^{(0)}(t),J_i]_- \exp\left[-i\frac{H^{(0)}t}{\hbar}\right]. \quad (137)$$

Using Eq. (136) for $J_i^{(1)}(t)$, the contribution $\langle\frac{1}{2}[J_i^{(1)}(t), J_j^{(1)}(t+\tau)]_+\rangle^{(0)}$ to the second-order steady-state current noise becomes

$$\langle\tfrac{1}{2}[J_i^{(1)},J_j^{(1)}(\tau)]_+\rangle^{(0)} =\sum_{kl}\mathcal{E}_k\mathcal{E}_l\langle\tfrac{1}{2}[\sigma_{ki},\sigma_{lj}^{(0)}(\tau)]_+\rangle^{(0)} \quad (138)$$

where we have set $t=0$ in virtue of time stationarity. The quantity $\langle\frac{1}{2}[\sigma_{ki},\sigma_{lj}^{(0)}(\tau)]_+\rangle^{(0)}$ can be interpreted as the second correlation moment between the spontaneous equilibrium fluctuations of the conductivity matrix elements σ_{ki} and $\sigma_{jl}^{(0)}(\tau)$.

The correlation function $\langle\frac{1}{2}[\sigma,\sigma^{(0)}(t)]_+\rangle^{(0)}$ is easily calculated in the case of a simple semiconductor for which the conductivity is given by $\sigma=ne^2\tau_s/m^*$,[15] where n is the equilibrium carrier concentration, m^* is the effective mass, and τ_s denotes a simple relaxation time for the scattering mechanisms. Assuming that

$$n(t)=n_0e^{-t/\tau_\beta} \quad (139)$$

where τ_β denotes a relaxation time associated with thermal charge carrier generation and recombination, the second-order term $\langle J^{(1)}J^{(1)}(t)\rangle^{(0)}$ in the steady-state current noise $\langle JJ(t)\rangle$ is

$$\langle J^{(1)}J^{(1)}(t)\rangle^{(0)}=\frac{e^4{\tau_s}^2}{m^{*2}}\langle nn(t)\rangle^{(0)}\mathcal{E}^2 =\frac{e^4{\tau_s}^2}{m^{*2}}\langle n^2\rangle^{(0)}e^{-t/\tau_\beta}\mathcal{E}^2 \quad (140)$$

which is a well-known result.[12]

[15] See for example W. Shockley, *Holes and Electrons in Semiconductors* (D. Van Nostrand Company, Inc., Princeton, New Jersey, 1950).

There exist in general other contributions to the second-order steady-state noise, as can be seen from Eq. (133). Here we have sought only to establish the connection of the general formalism with the usual model treatment of semiconductor noise.

13. HIGHER-ORDER STEP-DRIVEN RESPONSE

Whereas we have previously shown that the second-order response in a general process is characterized by the *difference* of two equilibrium third moments, the classical result of Eq. (38) suggests the possibility of establishing a more conventional quantum relationship in the case of step-driven processes. Although measurement of two distinct equilibrium third moments is still required to determine the second-order step-driven response, we shall find that for this simple class of processes the relationship is in close formal analogy to the first-order fluctuation-dissipation theorem. Further, the uniquely quantum-mechanical effects are more easily visualized in this case.

It is convenient to consider $\langle Q_i(t)\rangle^{(2)}$ as given in the form of Eq. (33). Rewriting this expression so as to indicate explicitly both contributions from a given pair of forces F_j, F_k, and assuming that $\langle Q_i\rangle^{(0)}=0$,

$$\langle Q_i(t)\rangle^{(2)}=\tfrac{1}{2}\sum_{jk}F_jF_k\int_0^\beta d\lambda_1\int_0^{\lambda_1}d\lambda_2 \times[\langle Q_j^{(0)}(-i\hbar\lambda_1)Q_k^{(0)}(-i\hbar\lambda_2)Q_i^{(0)}(t)\rangle^{(0)} +\langle Q_k^{(0)}(-i\hbar\lambda_1)Q_j^{(0)}(-i\hbar\lambda_2)Q_i^{(0)}(t)\rangle^{(0)}]. \quad (141)$$

Inserting $\exp[\pm\beta H^{(0)}]$ in front of the operator $Q_i^{(0)}(t)$ and permuting the operators cyclicly, the integrand of the second term in Eq. (141) can be rewritten

$$\langle Q_k^{(0)}(-i\hbar\lambda_1)Q_j^{(0)}(-i\hbar\lambda_2)Q_i^{(0)}(t)\rangle^{(0)} =\langle Q_i^{(0)}(t)Q_k^{(0)}(-i\hbar\lambda_1+i\hbar\beta)Q_j^{(0)} \times(-i\hbar\lambda_2+i\hbar\beta)\rangle^{(0)}. \quad (142)$$

Thus, inverting the order of integration, and making the successive transformations $\lambda_2'=\beta-\lambda_1$, $\lambda_1'=\beta-\lambda_2$,

$$\int_0^\beta d\lambda_1\int_0^{\lambda_1}d\lambda_2\langle Q_k^{(0)}(-i\hbar\lambda_1)Q_j^{(0)}(-i\hbar\lambda_2)Q_i^{(0)}(t)\rangle^{(0)} =\int_0^\beta d\lambda_1\int_0^{\lambda_1}d\lambda_2\langle Q_i^{(0)}(t)Q_k^{(0)}(+i\hbar\lambda_2) \times Q_j^{(0)}(+i\hbar\lambda_1)\rangle^{(0)}. \quad (143)$$

Inserting (143) into Eq. (141), we thereby obtain $\langle Q_i(t)\rangle^{(2)}$ in the form

$$\langle Q_i(t)\rangle^{(2)}=\tfrac{1}{2}\sum_{jk}F_jF_k\int_0^\beta d\lambda_1\int_0^{\lambda_1}d\lambda_2 \times[\langle Q_j^{(0)}(-i\hbar\lambda_1)Q_k^{(0)}(-i\hbar\lambda_2)Q_i^{(0)}(t)\rangle^{(0)} +\langle Q_i^{(0)}(t)Q_k^{(0)}(i\hbar\lambda_2)Q_j^{(0)}(i\hbar\lambda_1)\rangle^{(0)}]. \quad (144)$$

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

It is further convenient to define the second-order step-response function $\Phi_{jki}^{(2)}(t)$ by rewriting Eq. (144) as

$$\langle Q_i(t)\rangle^{(2)}=\tfrac{1}{2}\sum_{jk}F_jF_k\Phi_{jki}^{(2)}(t)$$

whence

$$\Phi_{jki}^{(2)}(t)=\int_0^\beta d\lambda_1\int_0^{\lambda_1}d\lambda_2[\langle Q_j^{(0)}(-i\hbar\lambda_1)\times Q_k^{(0)}(-i\hbar\lambda_2)Q_i^{(0)}(t)\rangle^{(0)}+(\text{c.c.})]. \quad (146)$$

The notation (c.c.) is used to indicate the complex conjugate of the first term. $\Phi_{jki}^{(2)}(t)$ represents the second-order response to unit step-function forces F_j and F_k.

We now proceed to analyze $\Phi_{jki}^{(2)}(t)$ by decomposing the equilibrium expectation values appearing in Eq. (146) into appropriate summations over matrix elements in the unperturbed energy representation. This technique is similar to that employed in Sec. 5 in connection with $\phi_{ij}^{(1)}(t)$, and yields

$$\Phi_{jki}^{(2)}(t)=\int_0^\beta d\lambda_1\int_0^{\lambda_1}d\lambda_2\sum_{lmn}\rho(E_l)\langle E_l|Q_j|E_m\rangle\times\langle E_m|Q_k|E_n\rangle\langle E_n|Q_i|E_l\rangle e^{\lambda_1(E_l-E_m)}\times e^{\lambda_2(E_m-E_n)}\exp\left[i\frac{(E_n-E_l)t}{\hbar}\right]+(\text{c.c.}) \quad (147)$$

where $\langle E_l|Q_j|E_m\rangle$ is the matrix element of Q_j between the eigenstates of $H^{(0)}$ having the energy eigenvalues E_l and E_m, and $\rho(E_l)=e^{-\beta E_l}/\sum_l e^{-\beta E_l}$.

Performing the indicated temperature integration, and replacing the triple summation by a triple integral over energy eigenvalues, this becomes

$$\Phi_{jki}^{(2)}(t)=\int_{-\infty}^{\infty}dE_l\int_{-\infty}^{\infty}dE_m\int_{-\infty}^{\infty}dE_n\rho(E_l)\eta(E_l)\times\eta(E_m)\eta(E_n)\left[\frac{1}{(E_l-E_n)(E_l-E_m)}-\frac{e^{\beta(E_l-E_m)}}{(E_m-E_n)(E_l-E_m)}+\frac{e^{\beta(E_l-E_n)}}{(E_m-E_n)(E_l-E_n)}\right]\times\langle E_l|Q_j|E_m\rangle\langle E_m|Q_k|E_n\rangle\langle E_n|Q_i|E_l\rangle\times\exp\left[i\frac{(E_n-E_l)t}{\hbar}\right]+(\text{c.c.}) \quad (148)$$

where $\eta(E_l)$ is the energy density-of-states function. Introducing the transformations

$$E_m=E_l+\hbar\omega_1,\quad E_n=E_l+\hbar\omega_2 \quad (149)$$

and letting $\omega_1\to-\omega_1$, $\omega_2\to-\omega_2$ in the (c.c.) term, we finally obtain $\Phi_{jki}^{(2)}(t)$ in the form

$$\Phi_{jki}^{(2)}(t)=\frac{1}{2\pi}\int_{-\infty}^{\infty}d\omega_1\int_{-\infty}^{\infty}d\omega_2e^{i\omega_2t}N_{jki}^{(2)}(\omega_1,\omega_2) \quad (150)$$

where

$$N_{jki}^{(2)}(\omega_1,\omega_2)=\left[\frac{1}{\omega_1\omega_2}+\frac{\exp(-\hbar\beta\omega_1)}{\omega_1(\omega_1-\omega_2)}+\frac{\exp(-\hbar\beta\omega_2)}{\omega_2(\omega_2-\omega_1)}\right]\times g_{jki}(\omega_1,\omega_2)+\left[\frac{1}{\omega_1\omega_2}+\frac{\exp(+\hbar\beta\omega_1)}{\omega_1(\omega_1-\omega_2)}+\frac{\exp(+\hbar\beta\omega_2)}{\omega_2(\omega_2-\omega_1)}\right]g_{jki}^*(-\omega_1,-\omega_2) \quad (151)$$

and

$$g_{jki}(\omega_1,\omega_2)=g_{ikj}^*(\omega_2,\omega_1)=2\pi\int_{-\infty}^{\infty}dE\rho(E)\eta(E)\eta(E+\hbar\omega_1)\times\eta(E+\hbar\omega_2)\langle E|Q_j|E+\hbar\omega_1\rangle\times\langle E+\hbar\omega_1|Q_k|E+\hbar\omega_2\rangle\langle E+\hbar\omega_2|Q_i|E\rangle. \quad (152)$$

In order to relate the second-order step-driven response to the equilibrium fluctuations, we undertake a similar spectral analysis of the equilibrium correlation moment of the variables corresponding to the operators Q_j, $Q_k^{(0)}(t_1)$, and $Q_i^{(0)}(t_1+t_2)$. However, as discussed at the end of Sec. 9, there exist several equally valid, operationally distinct quantum-mechanical expressions for a given equilibrium third moment. Thus, the quantity $\Psi_{jki}^{(0)}$ defined in Eq. (112) corresponds to one particular set of experimental conditions, while the fully symmetrized equilibrium form $\bar{\Psi}_{jki}^{(0)}$ defined in Eq. (113) is appropriate to a different experimental arrangement. We find, in fact, that measurement of both $\Psi_{jki}^{(0)}(t_1,t_2)$ and $\bar{\Psi}_{jki}^{(0)}(t_1,t_2)$ is required for a complete experimental determination of $\Phi_{jki}^{(2)}(t)$, except in the classical limit.

Equation (113) for $\bar{\Psi}_{jki}^{(0)}(t_1,t_2)$ is conveniently rewritten

$$\bar{\Psi}_{jki}^{(0)}(t_1,t_2)=\tfrac{1}{6}\langle Q_jQ_k^{(0)}(t_1)Q_i^{(0)}(t_1+t_2)+Q_j^{(0)}(-i\hbar\beta)Q_k^{(0)}(t_1)Q_i^{(0)}(t_1+t_2)+Q_j^{(0)}(-i\hbar\beta)Q_k^{(0)}(t_1-i\hbar\beta)\times Q_i^{(0)}(t_1+t_2)\rangle^{(0)}+(\text{c.c.}) \quad (153)$$

where we have inserted $\exp[\pm\beta H^{(0)}]$ at suitable positions in the second and third terms and permuted the operators cyclicly.

As in the case of $\Phi_{jki}^{(2)}(t)$, the quantity $\bar{\Psi}_{jki}^{(0)}(t_1,t_2)$ can be decomposed into a triple integral over matrix elements in the unperturbed energy representation. Introducing the transformations (149) and letting

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

$\omega_1 \rightarrow -\omega_1$, $\omega_2 \rightarrow -\omega_2$ in the (c.c.) term, we obtain the expression

$$\Psi_{jki}^{(0)}(t_1,t_2)=\frac{1}{2\pi}\int_{-\infty}^{\infty}d\omega_1\int_{-\infty}^{\infty}d\omega_2 e^{i\omega_1 t_1}e^{i\omega_2 t_2} \times\bar{G}_{jki}^{(0)}(\omega_1,\omega_2) \quad (154)$$

where

$$\bar{G}_{jki}^{(0)}(\omega_1,\omega_2) =\frac{h^2}{6}\{[1+\exp(-h\beta\omega_1)+\exp(-h\beta\omega_2)] \times g_{jki}(\omega_1,\omega_2)+[1+\exp(+h\beta\omega_1) +\exp(+h\beta\omega_2)]g_{jki}^*(-\omega_1,-\omega_2)\}, \quad (155)$$

$g_{jki}(\omega_1,\omega_2)$ having been given previously in Eq. (152). $\bar{G}_{jki}^{(0)}(\omega_1,\omega_2)$ is the double Fourier transform of the equilibrium third moment $\bar{\Psi}_{jki}^{(0)}(t_1,t_2)$.

Similar analysis of the equilibrium third moment $\Psi_{jki}^{(0)}(t_1,t_2)$ of Eq. (112) yields the result

$$\Psi_{jki}^{(0)}(t_1,t_2)=\frac{1}{2\pi}\int_{-\infty}^{\infty}d\omega_1\int_{-\infty}^{\infty}d\omega_2 e^{i\omega_1 t_1}e^{i\omega_2 t_2} \times G_{jki}^{(0)}(\omega_1,\omega_2) \quad (156)$$

where

$$G_{jki}^{(0)}(\omega_1,\omega_2) =\frac{h^2}{4}\{[1+\exp(-h\beta\omega_1)]g_{jki}(\omega_1,\omega_2) +[1+\exp(+h\beta\omega_1)]g_{jki}^*(-\omega_1,-\omega_2)\}. \quad (157)$$

We now discuss the relation of the second-order step-driven response to the equilibrium fluctuations using the spectral quantities $N_{jki}^{(2)}(\omega_1,\omega_2)$, $\bar{G}_{jki}^{(0)}(\omega_1,\omega_2)$, and $G_{jki}^{(0)}(\omega_1,\omega_2)$ obtained above. In order to indicate the formal analogy of this relationship to the first-order fluctuation-dissipation theorem, it is convenient to define the functions $r_{jki}^{(2)}(\omega_1,\omega_2)$ and $f_{jki}^{(0)}(\omega_1,\omega_2)$.

$$r_{jki}^{(2)}(\omega_1,\omega_2)=\left[\frac{1}{\omega_1\omega_2}+\frac{\exp(-h\beta\omega_1)}{\omega_1(\omega_1-\omega_2)}+\frac{\exp(-h\beta\omega_2)}{\omega_2(\omega_2-\omega_1)}\right] \times g_{jki}(\omega_1,\omega_2), \quad (158)$$

$$f_{jki}^{(0)}(\omega_1,\omega_2)=\frac{h^2}{6}[1+\exp(-h\beta\omega_1)+\exp(-h\beta\omega_2)] \times g_{jki}(\omega_1,\omega_2), \quad (159)$$

in terms of which Eqs. (159) and (155) can be rewritten

$$N_{jki}^{(2)}(\omega_1,\omega_2)=r_{jki}^{(2)}(\omega_1,\omega_2)+r_{jki}^{(2)*}(-\omega_1,-\omega_2), \quad (160)$$

$$\bar{G}_{jki}^{(0)}(\omega_1,\omega_2)=f_{jki}^{(0)}(\omega_1,\omega_2)+f_{jki}^{(0)*}(-\omega_1,-\omega_2). \quad (161)$$

Comparison of Eqs. (158) and (159) shows that $r_{jki}^{(2)}(\omega_1,\omega_2)$ and $f_{jki}^{(0)}(\omega_1,\omega_2)$, which characterize the second-order step-driven response and the equilibrium fluctuations, respectively, in the sense of Eqs. (160) and (161), are related according to

$$r_{jki}^{(2)}(\omega_1,\omega_2)=\frac{1}{E^{(2)}(\omega_1,\omega_2;\beta)}f_{jki}^{(0)}(\omega_1,\omega_2) \quad (162)$$

where

$$E^{(2)}(\omega_1,\omega_2;\beta) =\frac{h^2}{6}\frac{1+\exp(-h\beta\omega_1)+\exp(-h\beta\omega_2)}{\dfrac{1}{\omega_1\omega_2}+\dfrac{\exp(-h\beta\omega_1)}{\omega_1(\omega_1-\omega_2)}+\dfrac{\exp(-h\beta\omega_2)}{\omega_2(\omega_2-\omega_1)}} \underset{\beta\rightarrow 0}{\rightarrow}\frac{1}{\beta^2}. \quad (163)$$

Equation (162) strongly suggests itself as a direct extension to second-order processes of the fluctuation-dissipation theorem of Eq. (58), which relates the Fourier transform $L_{ij}^{(1)}(\omega)$ of the first-order aftereffect function $\phi_{ij}^{(1)}(t)$ to the Fourier transform $G_{ij}^{(0)}(\omega)$ of the equilibrium second fluctuation moment. The universal function $E^{(2)}(\omega_1,\omega_2;\beta)$ is the second-order analog of $E^{(1)}(\omega;\beta)$. It is uniquely quantum-mechanical in origin, corresponding to a slight smearing out of the microscopic contributions to the macroscopic response at extremely high frequencies. In the classical limit, $E^{(2)}(\omega_1,\omega_2;\beta)\rightarrow(1/\beta^2)$ as indicated.

Whereas Eq. (58) constitutes a true thermodynamic relationship, however, Eq. (162) does not. That is, the function $f_{jki}^{(0)}(\omega_1,\omega_2)$ is not macroscopically observable, although its sum with $f_{jki}^{*(0)}(-\omega_1,-\omega_2)$ is, as indicated by Eq. (161).

The specific way in which quantum-mechanical effects enter the picture is evident by first considering what happens in the classical limit. For this case, Eqs. (151), (155), and (157) reduce to

$$N_{jki}^{(2)}(\omega_1,\omega_2)=\frac{h^2\beta^2}{6}[g_{jki}(\omega_1,\omega_2)+g_{jki}^*(-\omega_1,-\omega_2)] =\beta^2\bar{G}_{jki}^{(0)}(\omega_1,\omega_2)=\beta^2 G_{jki}^{(0)}(\omega_1,\omega_2). \quad (164)$$

The functions $\bar{G}_{jki}^{(0)}(\omega_1,\omega_2)$ and $G_{jki}^{(0)}(\omega_1,\omega_2)$ have become equivalent, and measurement of either completely determines $N_{jki}^{(2)}(\omega_1,\omega_2)$. The temporal form of Eq. (164) is obtained immediately from Eqs. (150), (154), and (156).

$$\Phi_{jki}^{(2)}(t)\underset{\beta\rightarrow 0}{\rightarrow}\beta^2\langle q_j q_k q_i(t)\rangle^{(0)}. \quad (165)$$

Substitution of this result into Eq. (145) for $\langle Q_i(t)\rangle^{(2)}$ yields the classical result given previously in Eq. (38).

We decompose $g_{jki}(\omega_1,\omega_2)$ into its real and imaginary, even and odd parts with respect to simultaneous reversal of ω_1 and ω_2.

$$g_{jki}(\omega_1,\omega_2) =\mathrm{Re}^{(+)}g_{jki}(\omega_1,\omega_2)+i\,\mathrm{Im}^{(+)}g_{jki}(\omega_1,\omega_2) +\mathrm{Re}^{(-)}g_{jki}(\omega_1,\omega_2)+i\,\mathrm{Im}^{(-)}g_{jki}(\omega_1,\omega_2). \quad (166)$$

The superscripts (+) and (−) denote the even and

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

odd part, respectively, under the transformation $\omega_1 \rightarrow -\omega_1$, $\omega_2 \rightarrow -\omega_2$. Using this decomposition, the classical relationship (164) becomes

$$N_{jki}^{(2)}(\omega_1,\omega_2)=\frac{\hbar^2\beta^2}{3}[\mathrm{Re}^{(+)}g_{jki}(\omega_1,\omega_2) + i\,\mathrm{Im}^{(-)}g_{jki}(\omega_1,\omega_2)]=\beta^2\bar{G}_{jki}^{(0)}(\omega_1,\omega_2). \quad (167)$$

Thus, in the classical limit we are concerned only with $\mathrm{Re}^{(+)}g_{jki}(\omega_1,\omega_2)$ and $\mathrm{Im}^{(-1)}g_{jki}(\omega_1,\omega_2)$, both of which are determined by experimental knowledge of the complex quantity $\bar{G}_{jki}^{(0)}(\omega_1,\omega_2)=G_{jki}^{(0)}(\omega_1,\omega_2)$.

In the general quantum case, however, $N_{jki}^{(2)}(\omega_1,\omega_2)$ depends upon all four components of $q_{jki}(\omega_1,\omega_2)$ because of the uniquely quantum-mechanical spreading introduced by the quantities

$$\left[\frac{1}{\omega_1\omega_2}+\frac{\exp(\pm\hbar\beta\omega_1)}{\omega_1(\omega_1-\omega_2)}+\frac{\exp(\pm\hbar\beta\omega_2)}{\omega_2(\omega_2-\omega_1)}\right]$$

appearing in Eq. (151). For this case $\bar{G}_{jki}^{(0)}(\omega_1,\omega_2)$ and $G_{jki}^{(0)}(\omega_1,\omega_2)$ are no longer equivalent, and Eq. (155) and (157) constitute two independent complex expressions which can be solved for the four components of $g_{jki}(\omega_1,\omega_2)$. Because of the quantum-mechanical interference among the components of $g_{jki}(\omega_1,\omega_2)$, measurement of both equilibrium third moments $\bar{\Psi}_{jki}^{(0)}(t_1,t_2)$ and $\Psi_{jki}^{(0)}(t_1,t_2)$ is necessary to determine completely the function $N_{jki}^{(2)}(\omega_1,\omega_2)$ and, consequently, by Eq. (150), the second-order step-driven response, $\Phi_{jki}^{(2)}(t)$.

It may be presumed that an analogous quantum-mechanical analysis of the third- and higher-order terms in the step-driven response can be made, although we do not attempt to carry out this laborious program here. Instead we simply refer to the classical relation between the step-driven response and the equilibrium fluctuations, given previously in Eqs. (36) through (39).

14. STEP-DRIVEN NOISE

For step-driven processes, the perturbation expansion of the driven noise $\langle\frac{1}{2}[Q_i(t),Q_j(t+\tau)]_+\rangle$ simplifies considerably, exhibiting a strong formal similarity to the step-driven response $\langle Q_i(t)\rangle$. We can, therefore, use the techniques employed previously for analyzing the step-driven response to discuss the relationship between the step-driven noise and the equilibrium fluctuations.

Consider Eq. (117) for the first-order driven noise $\langle\frac{1}{2}[Q_i(t), Q_j(t+\tau)]_+\rangle^{(1)}$. For step-function forces, the unsymmetrized term vanishes, and Eq. (117) reduces, upon integration, to

$$\langle\tfrac{1}{2}[Q_i(t), Q_j(t+\tau)]_+\rangle^{(1)} = -\sum_k F_k\int_{-\infty}^{\infty} dt_1\Gamma(t_1)\langle\tfrac{1}{4}[Q_k^{(0)}(t_1), \times[Q_i^{(0)}(t), Q_j^{(0)}(t+\tau)]_+]_+\rangle^{(0)}. \quad (168)$$

Equation (168) can also be obtained from Eq. (88) for the first-order step-driven response $\langle Q_i(t)\rangle^{(1)}$ by replacing the operator Q_j by the anticommutator $\frac{1}{2}[Q_i,Q_j^{(0)}(\tau)]_+$. Equation (168) relates the first-order step-driven noise $\langle\frac{1}{2}[Q_i(t), Q_j(t+\tau)]_+\rangle^{(1)}$ to the third equilibrium correlation moment

$$\langle\tfrac{1}{4}[Q_k^{(0)}(t_1), [Q_i^{(0)}(t), Q_j^{(0)}(t+\tau)]_+]_+\rangle^{(0)}.$$

In the classical limit it reduces to

$$\langle q_i(t)q_j(t+\tau)\rangle^{(1)}=-\beta\sum_k F_k\langle q_kq_i(t)q_j(t+\tau)\rangle^{(0)}. \quad (169)$$

The formal similarity between Eqs. (88) and (168) also carries over into the higher-order terms. Thus, in the case of step-function forces, Eq. (21) for $\langle\frac{1}{2}[Q_i(t,) Q_j(t+\tau)]_+\rangle^{(2)}$ reduces to

$$\langle\tfrac{1}{2}[Q_i(t), Q_j(t+\tau)]_+\rangle^{(2)} = \left(-\frac{1}{i\hbar}\right)^2\sum_{kl}F_kF_l\int_{-\infty}^{0}dt_1\int_{-\infty}^{t_1}dt_2\langle[Q_l^{(0)}(t_2), \times[Q_k^{(0)}(t_1), \tfrac{1}{2}[Q_i^{(0)}(t), Q_j^{(0)}(t+\tau)]_+]_-]_-\rangle^{(0)} \quad (170)$$

which also follows from Eq. (24) for the second-order step-driven response $\langle Q_i(t)\rangle^{(2)}$, if we replace Q_i by $\frac{1}{2}[Q_i,Q_j^{(0)}(\tau)]_+$. Consequently, the analysis of the previous section for the second-order step-driven response can be applied directly to the second-order step-driven noise. Similarly, it may be assumed that the third- and higher-order terms can also be treated on an equivalent basis.

The consequences of the above formal resemblance can be most simply demonstrated in the classical limit. Consider Eq. (36) for the full step-driven response $\langle q_i(t)\rangle$. Replacing the operator Q_i by the anticommutator $\frac{1}{2}[Q_i,Q_j^{(0)}(\tau)]_+$, which corresponds to replacing the classical variable q_i by the product $q_iq_j(\tau)$, Eq. (36) becomes

$$\langle q_i(t)q_j(t+\tau)\rangle=\langle\exp(-\beta\sum_k F_kq_k)q_i(t)q_j(t+\tau)\rangle^{(0)}/\langle\exp(-\beta\sum_k F_kq_k)\rangle^{(0)}, \quad (171)$$

Equation (171) expresses the step-driven noise $\langle q_i(t)q_j(t+\tau)\rangle$ wholly in terms of the third correlation moment $\langle\exp(-\beta\sum_k F_kq_k)q_i(t)q_j(t+\tau)\rangle^{(0)}$ of the spontaneous equilibrium fluctuations.

It is possible to summarize the results of this section and the preceding one in the following intuitively appealing way. We first differentiate Eq. (36) for $\langle q_i(t)\rangle$ with respect to F_j, evaluating the result at F_1, F_2, $\cdots=0$.

$$\frac{\partial}{\partial F_j}\langle q_i(t)\rangle=-\beta\langle q_jq_i(t)\rangle^{(0)} \quad (172)$$

which is equivalent to Eq. (37).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

Taking the second derivative of $\langle q_i(t)\rangle$ with respect to F_j and F_k, we obtain

$$\frac{\partial^2}{\partial F_j \partial F_k}\langle q_i(t)\rangle = \beta^2\langle q_j q_k q_i(t)\rangle^{(0)} \tag{173}$$

which is equivalent to Eq. (38).

On the other hand, differentiating Eq. (171) for $\langle q_j(t)q_i(t+\tau)\rangle$ with respect to F_k, we obtain

$$\frac{\partial}{\partial F_k}\langle q_j(t)q_i(t+\tau)\rangle = -\beta\langle q_k q_j(t)q_i(t+\tau)\rangle^{(0)} \tag{174}$$

which is equivalent to Eq. (169). Letting $t=0$ and replacing τ by t, this becomes

$$\frac{\partial}{\partial F_k}\langle q_j q_i(t)\rangle = -\beta\langle q_k q_j q_i(t)\rangle^{(0)}. \tag{175}$$

Equations (173) and (175) can be combined in the form

$$\frac{\partial^2}{\partial F_j \partial F_k}\langle q_i(t)\rangle = -\beta\frac{\partial}{\partial F_k}\langle q_j q_i(t)\rangle = \beta^2\langle q_j q_k q_i(t)\rangle^{(0)} \tag{176}$$

which constitutes a triple relationship among the second-order response, the first-order noise, and the equilibrium third moment for a step-driven process. It is apparent that further differentiation of Eqs. (36) and (171) would yield a whole hierarchy of analogous higher order thermodynamic relationships.

Although they apply literally only to step-driven process, Eqs. (172) and (176) exhibit most of the essential elements of the more general theory. Therefore, the results presented above characterize the general structure of irreversible thermodynamics.

Finally, we compare our results for a step-driven process with the results of time-independent equilibrium fluctuation theory. Letting $t=0$, Eqs. (172) and (176) reduce to

$$\frac{\partial}{\partial F_j}\langle q_i\rangle^{(0)} = -\beta\langle q_i q_j\rangle^{(0)}, \tag{177}$$

$$\frac{\partial^2}{\partial F_j \partial F_k}\langle q_i\rangle^{(0)} = \beta^2\langle q_i q_j q_k\rangle^{(0)} = -\beta\frac{\partial}{\partial F_k}\langle q_i q_j\rangle^{(0)}. \tag{178}$$

Equations (177) and (178) are precisely those which can be derived for a generalized canonical ensemble using standard equilibrium fluctuation theory.[16]

15. THE PATH DISTRIBUTION FUNCTION

Recently one of us has discussed the first-order term in the nonequilibrium path distribution function $W_1(q,t)$[8]. This function specifies the probability that the macroscopic variable corresponding to the operator Q has the value q in the driven ensemble at the time t. In this section we review the theory of the path distribution function and its application to the first-order problem discussed in Secs. 5 through 7. We also discuss the path distribution function for a step-driven process.

In order to keep the notation simple, we restrict ourselves initially to a one-dimensional process. The distribution function $W_1(q,t)$ can be expressed in terms of its characteristic function $K_1(\nu,t)$.

$$W_1(q,t) = \frac{1}{(2\pi)^{\frac{1}{2}}}\int_{-\infty}^{\infty} d\nu e^{i\nu q}K_1(\nu,t) \tag{179}$$

whence $K_1(\nu,t)$ is given by

$$K_1(\nu,t) = \frac{1}{(2\pi)^{\frac{1}{2}}}\int_{-\infty}^{\infty} dq e^{-i\nu q}W_1(q,t)$$

$$= \frac{1}{(2\pi)^{\frac{1}{2}}}\langle e^{-i\nu q(t)}\rangle = \frac{1}{(2\pi)^{\frac{1}{2}}}\,\mathrm{Tr}\,\rho^{(0)}e^{-i\nu Q(t)}, \tag{180}$$

where $e^{-i\nu Q(t)} = U^\dagger(t)e^{-i\nu Q}U(t)$. Then

$$W_1(q,t) = \mathrm{Tr}\,\rho^{(0)}\delta(q-Q(t)) = \langle\delta(q-Q(t))\rangle, \tag{181}$$

where $\delta(q-Q(t))$ is the driven δ-function operator $U^\dagger(t)\delta(q-Q)U(t)$, which selects from $\rho^{(0)}$ the appropriate contributions to $W_1(q,t)$.

Since $W_1(q,t)$ is just the expectation of value of $\delta(q-Q)$ in the driven system at time t, the theory developed previously for the driven response $\langle Q(t)\rangle$ can be applied directly to the path distribution function. Thus, for example, the first-order term $W_1^{(1)}(q,t)$ is obtained in symmetrized form from Eq. (88) for $\langle Q_i(t)\rangle^{(1)}$ by simply replacing Q_i by $\delta(q-Q)$.

$$W_1^{(1)}(q,t) = -\int_{-\infty}^{t} dt_1 F(t_1)\int_{-\infty}^{\infty} dt_1'\Gamma(t_1-t_1')$$

$$\times\langle\tfrac{1}{2}[\dot{Q}^{(0)}(t_1'-t),\delta(q-Q)]_+\rangle^{(0)}, \tag{182}$$

the t dependence in the integrand has been transferred to $\dot{Q}$ by performing a unitary transformation with $\exp\{\pm i[H^{(0)}t/\hbar]$.

The quantity $\langle\frac{1}{2}[\dot{Q}^{(0)}(t_1'-t),\delta(q-Q)]_+\rangle^{(0)}$ is interpreted in the following way. Since the classical analog of the operator $\delta(q-Q)$ is simply the δ function $\delta(q-q')$, we can calculate the equilibrium correlation moment of the variables corresponding to the operators $\dot{Q}^{(0)}(t_1'-t)$ and $\delta(q-Q)$ according to

$$\langle\tfrac{1}{2}[\dot{Q}^{(0)}(t_1'-t),\delta(q-Q)]_+\rangle^{(0)}$$

$$= \int_{-\infty}^{\infty} dq'\delta(q-q')W_1^{(0)}(q')\langle\dot{Q}^{(0)}(t_1'-t)\rangle_{q'}^{(0)} \tag{183}$$

where $W_1^{(0)}(q)$ is the equilibrium probability distribu-

[16] See for example R. F. Greene and H. B. Callen, Phys. Rev. 83, 1231 (1951).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

tion, and $\langle\dot{Q}^{(0)}(t_1'-t)\rangle_q^{(0)}$ denotes the equilibrium expectation value of the variable corresponding to the operator $\dot{Q}$ at time $(t'-t)$ conditional on the variable corresponding to Q having the value q at time zero. The q' integration can be performed immediately to yield

$$\langle\tfrac{1}{2}[\dot{Q}^{(0)}(t_1'-t),\delta(q-Q)]_+\rangle^{(0)} = W_1^{(0)}(q)\langle\dot{Q}^{(0)}(t_1'-t)\rangle_q^{(0)}. \quad (184)$$

Substituting the result (184) back into Eq. (182) we obtain

$$W_1^{(1)}(q,t) = -W_1^{(0)}(q)\int_{-\infty}^{t} dt_1 F(t_1) \times\int_{-\infty}^{\infty} dt_1'\Gamma(t_1-t_1')\langle\dot{Q}^{(0)}(t_1'-t)\rangle_q^{(0)}. \quad (185)$$

The corresponding first-order term $W_1^{(1)}(q_1,\cdots,q_n;t)$ in the path distribution function for an n-dimensional process can be developed in a completely analogous way. In place of the single operator $\delta(q-Q)$, we introduce a symmetrized form of the product $\delta(q_1-Q_1)\cdots\delta(q_n-Q_n)$ of δ-function operators. The result is

$$W_1^{(1)}(q_1,\cdots,q_n;t) = -W_1^{(0)}(q_1,\cdots,q_n)\sum_{j=1}^{n}\int_{-\infty}^{t} dt_1 F_j(t_1) \times\int_{-\infty}^{\infty} dt_1'\Gamma(t_1-t_1')\langle\dot{Q}_j^{(0)}(t_1'-t)\rangle_{q_1\cdots q_n}^{(0)}. \quad (186)$$

$W_1^{(0)}(q_1,\cdots,q_n)$ is the simultaneous equilibrium probability distribution for the variables $q_1,\cdots,q_n$, while $\langle\dot{Q}_j^{(0)}(t_1'-t)\rangle^{(0)}{}_{q_1\cdots q_n}$ denotes the equilibrium expectation value of the variable corresponding to $\dot{Q}_j$ at time $(t_1'-t)$ conditional on the variables corresponding to $Q_i,\cdots,Q_n$ having the values $q_1,\cdots,q_n$ at time zero. In the classical limit, Eq. (186) reduces to

$$W_1^{(1)}(q_1,\cdots,q_n;t) \underset{\beta\to 0}{\longrightarrow} -\beta W_1^{(0)}(q_1,\cdots,q_n) \times\sum_{j=1}^{n}\int_{-\infty}^{t} dt_1 F_j(t_1)\langle\dot{q}_j(t_1-t)\rangle_{q_1\cdots q_n}^{(0)}. \quad (187)$$

Equation (186) constitutes a generalized statement of the fluctuation-dissipation theorem, expressing $W_1^{(1)}(q_1,\cdots,q_n;t)$ in terms of the equilibrium probability distribution $W_1^{(0)}(q_1,\cdots,q_n)$ and the equilibrium conditional expectation value $\langle\dot{Q}_j^{(0)}(t_1'-t)\rangle_{q_1\cdots q_n}^{(0)}$. Since all of the previous theorems regarding the first-order problem can be derived from Eq. (186), this form can be considered the fundamental relationship for the linear theory of irreversibility.

We now consider the full path distribution function $W_1(q_1,\cdots,q_n;t)$ for a classical step-driven process. The step-driven response is written in terms of $W_1(q_1,\cdots,q_n;t)$ as

$$\langle q_i(t)\rangle = \int_{-\infty}^{\infty} dq_1\cdots\int_{-\infty}^{\infty} dq_n q_i W_1(q_1,\cdots,q_n;t). \quad (188)$$

According to Eq. (36), however, $\langle q_i(t)\rangle$ can also be written

$$\langle q_i(t)\rangle = \int_{-\infty}^{\infty} dq_1\cdots\int_{-\infty}^{\infty} dq_n q_i W_1^{(0)}(q_1,\cdots,q_n) \times\langle\exp[-\beta\sum_j F_j q_j(-t)\rangle_{q_1\cdots q_n}^{(0)}/ \langle\exp(-\beta\sum_j F_j q_j)\rangle^{(0)}. \quad (189)$$

The t dependence of the time-stationary quantity $\langle\exp(-\beta\sum_j F_j q_j)q_i(t)\rangle^{(0)}$ has been translated into the exponential $\exp(-\beta\sum_j F_j q_j)$.

$$\langle\exp[-\beta\textstyle\sum_j F_j q_j(-t)]\rangle_{q_1\cdots q_n}^{(0)}$$

denotes the equilibrium expectation value of

$$\exp(-\beta\textstyle\sum_j F_j q_j)$$

at time $-t$ conditional on all variables having the values $q_1,\cdots,q_n$ at time zero.

Since Eqs. (188) and (189) must be identical, it follows that the step-driven path distribution function $W_1(q_1,\cdots,q_n;t)$ is given, in the classical limit, by

$$W_1(q_1,\cdots,q_n;t) = W_1^{(0)}(q_1,\cdots,q_n)\langle\exp[-\beta\sum_j F_j q_j(-t)]\rangle_{q_1\cdots q_n}^{(0)}/ \langle\exp-\beta\sum_j F_j q_j)\rangle^{(0)}. \quad (190)$$

Equation (190) expresses the path distribution function, characterizing the time evolution of a step-driven ensemble, in terms of the equilibrium probability distribution $W_1^{(0)}(q_1,\cdots,q_n)$, together with the conditional expectation value $\langle\exp[-\beta\sum_j F_j q_j(-t)\rangle_{q_1\cdots q_n}^{(0)}$.

The significance of this result for the path distribution function can be made more apparent by rewriting Eq. (36) for $\langle q_i(t)\rangle$ in the form

$$\langle q_i(t)\rangle = \int_{-\infty}^{\infty} dq_1\cdots\int_{-\infty}^{\infty} dq_n W_1^{(0)}(q_1,\cdots,q_n) \times\exp(-\beta\sum_j F_j q_j)\langle q_i(t)\rangle_{q_1\cdots q_n}^{(0)}/ \langle\exp(-\beta\sum_j F_j q_j)\rangle^{(0)}. \quad (191)$$

However, letting $t=0$ in Eq. (189), we find that the initial perturbed equilibrium probability distribution

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

$W_1(q_1,\cdots,q_n;0)$ is given by

$$W_1(q_1,\cdots q_n;0) = W_1^{(0)}(q_1;\cdots,q_n)\exp(-\beta\sum_j F_j q_j)/ \times\langle\exp(-\beta\sum_j F_j q_j)\rangle^{(0)}. \quad (192)$$

Inserting this result in Eq. (191), we obtain

$$\langle q_i(t)\rangle = \int_{-\infty}^{\infty} dq_1\cdots\int_{-\infty}^{\infty} dq_n W_1(q_1,\cdots,q_n;0) \times\langle q_i(t)\rangle_{q_1\cdots q_n}{}^{(0)}. \quad (193)$$

Equation (193) shows explicitly how the step-driven response $\langle q_i(t)\rangle$ is built up from the regression of the equilibrium fluctuations, characterized by the equilibrium conditional expectation value $\langle q_i(t)\rangle_{q_1\cdots q_n}{}^{(0)}$, and weighted according to the initial perturbed distribution $W_1(q_1,\cdots,q_n;0)$. Although this result is precisely that which we might intuitively expect, it has often been pointed out in the literature[2,6] that there is no clear *a priori* justification for identifying the behavior of a system undergoing an irreversible process with the spontaneous equilibrium fluctuations in this way. The equilibrium fluctuations are microscopic in nature and generally on an extremely small scale, whereas the macroscopic response functions measured in the laboratory are normally orders of magnitude larger. Nevertheless, the proof of the assumption that macroscopic processes follow the same laws of regression as the equilibrium fluctuations is provided by Eq. (193).

16. THE JOINT PATH DISTRIBUTION FUNCTION

Just as a more detailed description of the equilibrium behavior can be obtained by introducing joint probability distributions containing two or more times, it is possible to describe in greater detail the evolution of a driven ensemble by introducing joint path distribution functions. In this final section we extend the theory of the previous section to include the joint path distribution $W_2(q,t;q',t')$, which specifies the probability that the variable corresponding to the operator Q has the value q at time t *and* the value q' at time t' in a driven ensemble. We consider explicitly the case of a single variable, although the extension of the theory to multidimensional processes is quite straightforward.

$W_2(q,t;q',t')$ can be expressed in terms of its characteristic function $K_2(\nu,t;\nu',t')$.

$$W_2(q,t;q',t') = \frac{1}{2\pi}\int_{-\infty}^{\infty} d\nu\int_{-\infty}^{\infty} d\nu' e^{i\nu q}e^{i\nu' q'}K_2(\nu,t;\nu',t') \quad (194)$$

where $K_2(\nu,t;\nu',t')$ is given by

$$K_2(\nu,t;\nu',t') = \frac{1}{2\pi}\int_{-\infty}^{\infty} dq\int_{-\infty}^{\infty} dq' e^{-i\nu q}e^{-i\nu' q'}W_2(q,t;q',t') = \frac{1}{2\pi}\langle e^{-i\nu Q(t)}e^{-i\nu' Q(t')}\rangle. \quad (195)$$

According to the discussion of Sec. 3, the quantum-mechanical form of the driven second moment $K_2(\nu,t;\nu',t')$ is

$$K_2(\nu,t;\nu',t') = \frac{1}{2\pi}\langle\tfrac{1}{2}[e^{-i\nu Q(t)},e^{-i\nu' Q(t')}]_+\rangle^{(0)} \quad (196)$$

so that Eq. (194) for $W_2(q,t;q',t')$ becomes

$$W_2(q,t;q',t') = \frac{1}{(2\pi)^2}\int_{-\infty}^{\infty} d\nu\int_{-\infty}^{\infty} d\nu'\langle\tfrac{1}{2}[e^{i\nu[q-Q(t)]},e^{i\nu'[q'-Q(t')]}]_+\rangle^{(0)} = \langle\tfrac{1}{2}[\delta(q-Q(t)),\delta(q'-Q(t'))]_+\rangle^{(0)}. \quad (197)$$

Equation (197) states that $W_2(q,t;q',t')$ is just the driven correlation moment between the δ-function operators $\delta(q-Q(t))$ and $\delta(q'-Q(t'))$. Therefore, the treatment of driven second moments developed throughout the preceding sections of this paper is immediately applicable to the joint path distribution function.

We limit ourselves here to a discussion of the joint path distribution function for a classical step-driven process. The driven second moment $\langle q(t)q(t')\rangle$ can be written in terms of $W_2(q,t;q',t')$ as

$$\langle q(t)q(t')\rangle = \int_{-\infty}^{\infty} dq\int_{-\infty}^{\infty} dq' qq' W_2(q,t;q',t'). \quad (198)$$

On the other hand, according to Eq. (171), $\langle q(t)q(t')\rangle$ can be computed according to

$$\langle q(t)q(t')\rangle = \int_{-\infty}^{\infty} dq\int_{-\infty}^{\infty} dq' qq' W_2^{(0)}(q;q',t-t') \times\frac{\langle e^{-\beta F q(-t')}\rangle_{q,q'(t-t')}{}^{(0)}}{\langle e^{-\beta F q}\rangle^{(0)}} \quad (199)$$

where $W_2^{(0)}(q;q',t-t')$ is the (time-stationary) equilibrium joint probability distribution and $\langle e^{-\beta F q(-t')}\rangle_{q,q'(t-t')}{}^{(0)}$ denotes the equilibrium expectation value of $e^{-\beta F q}$ at time $-t'$ conditional on q at time zero and q' at time $(t-t')$.

From Eqs. (198) and (199) it follows that the step-driven joint path distribution function is given in the

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

classical limit, by

$$W_2(q,t;q',t')=W_2^{(0)}(q;q',t-t')\times\frac{\langle e^{-\beta Fq(-t')}\rangle_{q,q'(t-t')}^{(0)}}{\langle e^{-\beta Fq}\rangle^{(0)}}. \quad (200)$$

Thus, $W_2(q,t;q',t')$ is related in a particularly simple way to the equilibrium joint probability $W_2^{(0)}(q;q',t-t')$ and the equilibrium conditional expectation value $\langle e^{-\beta Fq(-t')}\rangle_{q,q'(t-t')}^{(0)}$.

Finally, we rewrite Eq. (171) for $\langle q(t)q(t')\rangle$ so as to further emphasize the relationship of this quantity to the regression of equilibrium fluctuations.

$$\langle q(t)q(t')\rangle=\int_{-\infty}^{\infty}dqW_1^{(0)}(q)e^{-\beta Fq}\frac{\langle q(t)q(t')\rangle_q^{(0)}}{\langle e^{-\beta Fq}\rangle^{(0)}}. \quad (201)$$

$\langle q(t)q(t')\rangle_q^{(0)}$ denotes the second equilibrium correlation moment between $q(t)$ and $q(t')$, conditional on the value q at time zero. However, according to Eq. (192) $W_1^{(0)}(q)e^{-\beta Fq}/\langle e^{-\beta Fq}\rangle^{(0)}$ is just the initial ($t=0$) perturbed equilibrium distribution function $W_1(q,0)$. Hence, Eq. (200) becomes

$$\langle q(t)q(t')\rangle=\int_{-\infty}^{\infty}dqW_1(q,0)\langle q(t)q(t')\rangle_q^{(0)}. \quad (202)$$

This result shows explicitly how the step-driven second moment $\langle q(t)q(t')\rangle$ is built up from the regression of the equilibrium fluctuations, characterized by the conditional equilibrium second moment $\langle q(t)q(t')\rangle_q^{(0)}$ and weighted according to the initial perturbed distribution $W_1(q,0)$.

APPENDIX A

In this appendix we compare a system driven from $t\to-\infty$ by the step-function forces defined in Sec. 4 to the subsequent motion of a system characterized at $t=0$ by the generalized canonical density operator $\rho(0)$ of Eq. (25).

Consider the first-order term $\langle Q_i(t)\rangle^{(1)}$ in the response during a step-driven process, as given by Eq. (23). The equilibrium expectation value appearing in this expression can be written

$$\begin{aligned}&\langle[Q_j^{(0)}(t_1),Q_i^{(0)}(t)]_-\rangle^{(0)}\\&=\langle Q_j^{(0)}(t_1)Q_i^{(0)}(t)-Q_i^{(0)}(t)\exp[\pm\beta H^{(0)}]Q_j^{(0)}(t_1)\rangle^{(0)}\\&=\langle Q_j^{(0)}(t_1)Q_i^{(0)}(t)-\exp[\beta H^{(0)}]Q_j^{(0)}(t_1)\\&\qquad\times\exp[-\beta H^{(0)}]Q_i^{(0)}(t)\rangle^{(0)}\end{aligned} \quad (A\text{-}1)$$

where we have inserted $\exp[\pm\beta H^{(0)}]$ in the second term as indicated and permuted the operators cyclicly.

This form can be further rewritten as follows.

$$\begin{aligned}&\langle[Q_j^{(0)}(t_1),Q_i^{(0)}(t)]_-\rangle^{(0)}\\&=-\int_0^\beta d\lambda_1\frac{\partial}{\partial\lambda_1}\langle\exp[\lambda_1H^{(0)}]Q_j^{(0)}(t_1)\\&\qquad\times\exp[-\lambda H^{(0)}]Q_i^{(0)}(t)\rangle^{(0)}\\&=-\int_0^\beta d\lambda_1\langle[H^{(0)},Q_j^{(0)}(t_1-i\hbar\lambda_1)]_-Q_i^{(0)}(t)\rangle^{(0)}.\end{aligned} \quad (A\text{-}2)$$

Noting that $[H^{(0)},Q_j^{(0)}(t_1-i\hbar\lambda_1)]=-i\hbar\dot{Q}_j^{(0)}(t_1-i\hbar\lambda_1)$, we have

$$\begin{aligned}&\langle[Q_j^{(0)}(t_1),Q_i^{(0)}(t)]_-\rangle^{(0)}\\&\qquad=i\hbar\int_0^\beta d\lambda_1\langle\dot{Q}_j^{(0)}(t_1-i\hbar\lambda_1)Q_i^{(0)}(t)\rangle^{(0)}.\end{aligned} \quad (A\text{-}3)$$

Inserting the result (A-3) into Eq. (23) and performing the time integration, we obtain

$$\begin{aligned}\langle Q_i(t)\rangle^{(1)}=-\sum_jF_j\int_0^\beta d\lambda_1[&\langle Q_j^{(0)}(-i\hbar\lambda_1)Q_i^{(0)}(t)\rangle^{(0)}\\&-\lim_{t_1\to-\infty}\langle Q_j^{(0)}(t_1-i\hbar\lambda_1)Q_i^{(0)}(t)\rangle^{(0)}].\end{aligned} \quad (A\text{-}4)$$

The contribution to (A-4) from the $t_1\to-\infty$ limit is evaluated by taking

$$\begin{aligned}&\lim_{t_1\to-\infty}\int_0^\beta d\lambda_1\langle Q_j^{(0)}(t_1-i\hbar\lambda_1)Q_i^{(0)}(t)\rangle^{(0)}\\&=\lim_{T\to-\infty}\frac{1}{T}\int_0^\beta d\lambda_1\int_0^T dt_1\langle Q_j^{(0)}(t_1-i\hbar\lambda_1)Q_i^{(0)}(t)\rangle^{(0)}.\end{aligned} \quad (A\text{-}5)$$

Because of the factor $1/T$, the oscillatory part of the integrand gives no contribution, and we are left with

$$\begin{aligned}&\lim_{t_1\to-\infty}\int_0^\beta d\lambda_1\langle Q_j^{(0)}(t_1-i\hbar\lambda_1)Q_i^{(0)}(t)\rangle^{(0)}\\&\qquad=\int_0^\beta d\lambda_1\langle\tilde{Q}_i\tilde{Q}_j\rangle^{(0)}=\beta\langle\tilde{Q}_i\tilde{Q}_j\rangle^{(0)}\end{aligned} \quad (A\text{-}6)$$

where $\tilde{Q}_1$ denotes the diagonal portion of the operator Q_i with respect to the unperturbed Hamiltonian $H^{(0)}$.

Using the result (A-6), Eq. (A-4) for $\langle Q_i(t)\rangle^{(1)}$ becomes

$$\begin{aligned}\langle Q_i(t)\rangle^{(1)}=-\sum_jF_j\Big[&\int_0^\beta d\lambda_1\langle Q_j^{(0)}(-i\hbar\lambda_1)Q_i^{(0)}(t)\rangle^{(0)}\\&-\beta\langle\tilde{Q}_i\tilde{Q}_j\rangle^{(0)}\Big].\end{aligned} \quad (A\text{-}7)$$

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

The paradoxical contrast between Eq. (A-7) and Eq. (32) for $\langle Q_i(t)\rangle^{(1)}$ has been discussed by R. Kubo,[9] who suggests that the former equation refers to an adiabatic system, whereas the latter refers to an isothermal system, so that the two need not be equal.

However, the same difficulty is found to arise even in an equilibrium isothermal system. Thus, in the limit $t\rightarrow\infty$, the equilibrium correlation moment

$$\langle \tfrac{1}{2}[Q_i,Q_j^{(0)}(t)]_+\rangle^{(0)}$$

becomes

$$\lim_{t\rightarrow\infty}\langle \tfrac{1}{2}[Q_i,Q_j^{(0)}(t)]_+\rangle^{(0)}=\langle \tilde{Q}_i\tilde{Q}_j\rangle^{(0)} \qquad \text{(A-8)}$$

where the limit is evaluated as in Eq. (A-6). In order for the ensemble to be ergodic, however, in the sense that as $t\rightarrow\infty$ the quantities involved become completely uncorrelated, we require

$$\lim_{t\rightarrow\infty}\langle \tfrac{1}{2}[Q_i,Q_j^{(0)}(t)]_+\rangle^{(0)}=\langle Q_i\rangle^{(0)}\langle Q_j\rangle^{(0)}. \qquad \text{(A-9)}$$

We believe that the resolution of this problem lies in the following interpretation. Throughout the discussion we consider an ensemble in continual interaction with a temperature reservoir. This should be contrasted with the interpretation adopted by Kubo, which is that the interaction with the temperature reservoir is removed at the moment of imposition of the applied forces, the ensemble thereafter being adiabatic. In our interpretation the Hamiltonian $H^{(0)}$ therefore contains a term corresponding to the interaction with a temperature reservoir, which we have not indicated explicitly for reasons to be explained momentarily. The additional interaction term induces incoherent transitions among the states of the system such that the ensemble "forgets" the details of its previous behavior after a sufficiently long time. This insures that the ensemble satisfies the ergodic requirement states in Eq. (A-9).

The justification for not indicating the interaction term explicitly in calculating the driven response is as follows. It is always possible to choose the term of interaction with the temperature reservoir to be so small that, for times comparable to those in which we are interested, the disordering effects arising from this source are negligible. Since Eq. (32) gives the first-order response $\langle Q_i(t)\rangle^{(1)}$ corresponding to an ensemble chosen so as to be in (generalized) canonical equilibrium up to $t=0$, it must therefore yield the appropriate evolution of $\langle Q_i(t)\rangle^{(1)}$ for any finite time $t>0$.

However, in the limit $t\rightarrow\infty$, the effects of the continued temperature interaction manifest themselves, regardless of the strength of the interaction. In this limit Eq. (32) reduces to

$$\langle Q_i(t)\rangle^{(1)}\underset{t\rightarrow\infty}{\rightarrow}-\sum_j F_j[\beta\langle \tilde{Q}_i\tilde{Q}_j\rangle^{(0)}-\beta\langle Q_i\rangle^{(0)}\langle Q_j\rangle^{(0)}]. \qquad \text{(A-10)}$$

Here it is necessary to take explicit account of the interaction with the temperature reservoir. This can be accomplished, according to a comparison of Eqs. (A-8) and (A-9), by replacing the quantity $\tilde{Q}_i$ by the average value $\langle Q_i\rangle^{(0)}$. The foregoing arguments apply as well to the time evolution of the full step-driven response $\langle Q_i(t)\rangle$ given in Eq. (26), in the expansion of which Eq. (32) is the first-order term.

In order to make Eq. (A-7) consistent with our interpretation, we recall that the quantity $\tilde{Q}_i$ appearing in it arose from the evaluation of a $t\rightarrow\infty$ limit. Again we take explicit account of the temperature interaction in this limit by replacing $\tilde{Q}_i$ by $\langle Q_i\rangle^{(0)}$. Thus, Eq. (A-7) becomes identical to Eq. (32). If we evaluate Eq. (24), for $\langle Q_i(t)\rangle^{(2)}$ and the corresponding higher-order terms in the response $\langle Q_i(t)\rangle$ during a step-driven process, replacing quantities of the type $\tilde{Q}_i$ by $\langle Q_i\rangle^{(0)}$ whenever they appear, we obtain expressions identical to Eqs. (33), (34), etc. The technique for accomplishing this is essentially an iteration of that employed in putting Eq. (23) into the form (A-7).

APPENDIX B

The causal nature of a linear process finds expression in the well-known Kramers-Krönig dispersion formulas relating the real and imaginary parts of the complex admittance matrix elements.

We first indicate the proof of these relations. Consider the Fourier transform of Eq. (63) for $Y_{ij}(\omega)$.

$$\int_{-\infty}^{\infty}d\omega e^{i\omega t}\frac{Y_{ij}(\omega)}{i\omega}=\int_{-\infty}^{\infty}d\omega\int_0^{\infty}dt'e^{i\omega(t-t')}\phi_{ij}^{(1)}(t')$$

$$=2\pi\int_0^{\infty}dt'\delta(t-t')\phi_{ij}^{(1)}(t'). \qquad \text{(B-1)}$$

Therefore,

$$\int_{-\infty}^{\infty}d\omega e^{i\omega t}\frac{Y_{ij}(\omega)}{i\omega}=\begin{cases}2\pi\phi_{ij}^{(1)}(t) & \text{for } t>0\\ 0 & \text{for } t<0\end{cases} \qquad \text{(B-2)}$$

whence it follows that $Y_{ij}(\omega)/i\omega$ can have poles only in the upper half of the complex ω plane.

For a function $Y_{ij}(\omega)/i\omega$ which is everywhere analytic in the lower half of the ω plane, Cauchy's integral theorem states that

$$\frac{Y_{ij}(\omega)}{i\omega}=\frac{P}{i\pi}\oint d\omega'\frac{Y_{ij}(\omega')}{i\omega'(\omega'-\omega)}$$

$$=-\frac{P}{i\pi}\int_{-\infty}^{\infty}d\omega'\frac{Y_{ij}(\omega')}{i\omega'(\omega'-\omega)} \qquad \text{(B-3)}$$

where the complex integration is taken around the contour shown in Fig. 5, and P denotes a Cauchy principal value. Decomposing $Y_{ij}(\omega)$ into its real and imaginary parts $Y_{ij}(\omega)=\mathrm{Re}\,Y_{ij}(\omega)+i\,\mathrm{Im}\,Y_{ij}(\omega)$ and equating the real and imaginary parts of Eq. (B-3), we

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

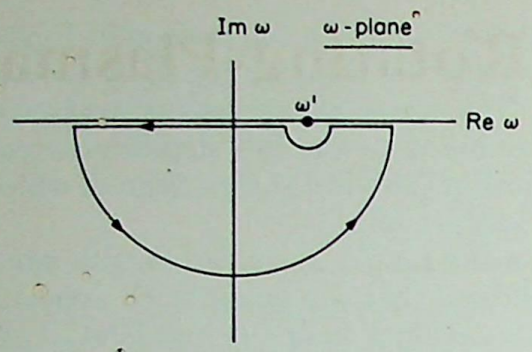


FIG. 5.

obtain

$$\frac{\text{Re } Y_{ij}(\omega)}{\omega}=-\frac{P}{\pi}\int_{-\infty}^{\infty}d\omega'\frac{\text{Im } Y_{ij}(\omega')}{\omega'(\omega'-\omega)}, \tag{B-4}$$

$$\frac{\text{Im } Y_{ij}(\omega)}{\omega}=\frac{P}{\pi}\int_{-\infty}^{\infty}d\omega'\frac{\text{Re } Y_{ij}(\omega')}{\omega'(\omega'-\omega)}. \tag{B-5}$$

Equations (B-4) and (B-5) are the Kramers-Krönig dispersion relations.

We use Eq. (B-5), together with the results of equilibrium fluctuation theory, to derive the $t=0$ form of the fluctuation-dissipation theorem, Eq. (73). Letting $\omega=0$, the quantity $[\text{Im } Y_{ij}(\omega)/\omega]$ becomes simply the capacitance $[\partial\langle q_j\rangle^{(0)}/\partial F_i]$, so that Eq. (B-5) assumes the form

$$\frac{\partial\langle q_j\rangle^{(0)}}{\partial F_i}=\frac{P}{\pi}\int_{-\infty}^{\infty}d\omega'\frac{\text{Re } Y_{ij}(\omega')}{\omega'^2}. \tag{B-6}$$

However, for a generalized canonical ensemble in contact with a series of reservoirs with constant intensive parameters $\cdots, F_i, F_j, \cdots$, the equilibrium second moment $\langle q_iq_j\rangle^{(0)}$ is given by[16]

$$\langle q_iq_j\rangle^{(0)}=-\frac{1}{\beta}\frac{\partial\langle q_j\rangle^{(0)}}{\partial F_i}. \tag{B-7}$$

Substitution of Eq. (B-7) into Eq. (B-6) yields the result

$$\langle q_iq_j\rangle^{(0)}=-\frac{P}{\pi\beta}\int_{-\infty}^{\infty}d\omega\frac{\text{Re } Y_{ij}(\omega)}{\omega^2} \tag{B-8}$$

which is identical to Eq. (73) with t set$=0$.

APPENDIX C

In this Appendix, following Kubo,[9] we evaluate the universal function $\Gamma(t)$, defined in Eq. (83). For $t>0$, this can be calculated by performing a contour integration around the upper half of the complex ω plane, while for $t<0$, the integration is taken around the lower half-plane. We consider explicitly the case $t>0$, since the $t<0$ calculation proceeds in an identical way.

Letting $\alpha=\hbar\beta\omega$, Eq. (83) can be rewritten as a contour integral in the complex α plane.

$$\Gamma(t)=\frac{1}{\pi\hbar}\oint d\alpha\exp\left(\frac{\alpha t}{i\hbar\beta}\right)\frac{1}{\alpha}\left(\frac{1-e^{-\alpha}}{1+e^{-\alpha}}\right). \tag{C-1}$$

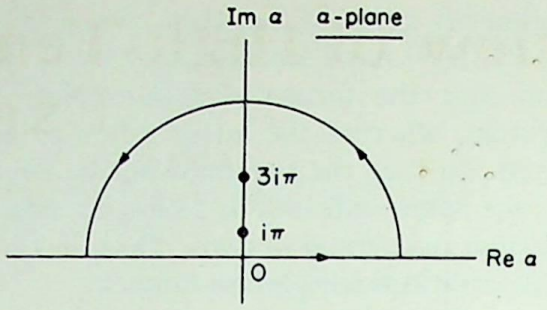


FIG. 6.

The integral is to be taken along the contour shown in Fig. 6. The integrand has simple poles at $\alpha_l=(2l+1)i\pi$, $l=0, \pm1, \cdots$. The residue of the integrand at the pole is obtained by letting $\alpha=\alpha_l+\Delta\alpha$, multiplying the integrand by $\Delta\alpha$, and taking the limit as $\Delta\alpha\to0$. Thus

residue at α_l

$$=\lim_{\Delta\alpha\to0}\frac{\Delta\alpha}{\pi\hbar}\exp\left[i\frac{(\alpha_l+\Delta\alpha)}{\hbar\beta}t\right]\frac{1}{(\alpha_l+\Delta\alpha)}\left[\frac{1-e^{-(\alpha_l+\Delta\alpha)}}{1+e^{-(\alpha_l+\Delta\alpha)}}\right]$$

$$=\frac{2}{\pi\hbar}\frac{\exp[-(2l+1)\pi t/\hbar\beta]}{(2l+1)\pi i}\lim_{\Delta\alpha\to0}\frac{\Delta\alpha}{(1-e^{-\Delta\alpha})}$$

$$=\frac{2\exp[-(2l+1)\pi t/\hbar\beta]}{\pi^2\hbar(2l+1)i}. \tag{C-2}$$

Equation (C-1) ls now evaluated by taking

$$\Gamma(t)=2\pi i\sum_{l=0}^{\infty}(\text{residue at }\alpha_l)$$

$$=\frac{4}{\pi\hbar}\sum_{l=0}^{\infty}\frac{\exp[-(2l+1)\pi t/\hbar\beta]}{(2l+1)}. \tag{C-3}$$

It is convenient to take the time derivative of Eq. (C-3) before performing the summation.

$$\frac{d\Gamma(t)}{dt}=-\frac{4}{\hbar^2\beta}\sum_{l=0}^{\infty}\exp[-(2l+1)\pi t/\hbar\beta]$$

$$=-\frac{4}{\hbar^2\beta}\frac{\exp[-(\pi t/\hbar\beta)]}{1-\exp(-2\pi t/\hbar\beta)}=-\frac{2}{\hbar^2\beta}\,\text{csch}\frac{\pi t}{\hbar\beta}. \tag{C-4}$$

In performing the summation in Eq. (C-4) we have made use of the expansion $\sum_l x^l=[1/(1-x)]$. We now integrate Eq. (C-4) with respect to t, which yields finally the desired result,

$$\Gamma(t)=\frac{2}{\pi\hbar}\ln\coth\frac{\pi t}{2\hbar\beta}. \tag{C-5}$$

The corresponding result for $t<0$ is identical to Eq. (C-5), except that t is replaced by $-t$. Thus, for all t, we can write that

$$\Gamma(t)=\frac{2}{\pi\hbar}\ln\coth\frac{\pi|t|}{2\hbar\beta}. \tag{C-6}$$

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

Review of High-Temperature Rotating-Plasma Experiments

JOHN M. WILCOX

Lawrence Radiation Laboratory, University of California, Berkeley, California

THE major experimental effort in the controlled thermonuclear reaction program has been based upon three different methods of plasma confinement—the pinch approach, the stellarator approach, and the magnetic-mirror approach.[1] Among the other approaches being investigated is the rotating-plasma system. By means of crossed electric and magnetic fields in a geometry having cylindrical symmetry, a charged particle can be induced to rotate about the axis of symmetry. The new degree of freedom associated with this rotation may offer new possibilities for containing and heating plasmas. Early work in this field was done at Oak Ridge.[2] We discuss here four experimental approaches that are being actively investigated: the Berkeley Homopolar device,[3] the Los Alamos Ixion,[4] and the Berkeley[5] and Moscow[6] ion-magnetron experiments.

BASIC CONCEPTS

We first discuss a few concepts which are important in these experiments, and then describe the four experiments individually. We will examine the electric drift velocity with the resulting diamagnetic current, the dielectric constant of a rotating plasma, the magnetic-mirror enhancement with a rotating plasma, and finally the possible formation of anode sheaths.

The well-known "electric drift velocity"[7] is imparted to a charged particle in the presence of crossed electric and magnetic fields. For a singly charged particle we have

$$\mathbf{V}_E = c(\mathbf{E}\times\mathbf{B})/B^2, \tag{1}$$

or

$$V_E = 10^8 E/B, \tag{2}$$

where the electric drift velocity V_E is in cm/sec, the electric field E is in volts/cm, and the magnetic field B is in gauss. Thus, for example, in an electric field of 3 kv/cm and a magnetic field of 10 kilogauss, a deuteron will have a velocity of 3×10^7 cm/sec and thus an energy of 1 kv. The electric drift velocity is perpendicular to both the electric and magnetic fields. It does not depend on the mass of the particle, and it is in the same direction for positive and negative particles. Thus, in an infinite plane geometry, ions and electrons have the same velocity and direction and there is no net electric current flow. However, in these experiments the presence of a radial electric field causes the particles to move with a radius of curvature R, and a centrifugal force term must be considered, so that the balance of forces on a particle is now

$$e[\mathbf{E}+(1/c)\mathbf{V}_D\times\mathbf{B}]+mV_D^2\mathbf{R}/R^2=0, \tag{3}$$

where $\mathbf{R}$ is the radius vector of the particle position, m is the mass of the particle, e is the algebraic value of its charge, and c is the velocity of light. This leads to a drift velocity

$$\mathbf{V}_D = c\frac{\mathbf{E}\times\mathbf{B}}{B^2} - \frac{mcV_D^2}{e}\frac{\mathbf{B}\times\mathbf{R}}{B^2R^2}. \tag{4}$$

Ions and electrons now move with different velocities, and an electric current $\mathbf{j}_\theta$ flows. Since the $\mathbf{j}_\theta\times\mathbf{B}$ force of the magnetic field is opposing the centrifugal force of the rotating particles, the plasma "leans" on the magnetic lines, and if the plasma rotational energy is sufficient the magnetic lines bow outward, thus weakening the field near the axis. The diamagnetic drift current density is

$$\mathbf{j}_\theta = (Ne/c)(\mathbf{V}_d - \mathbf{V}_e), \tag{5}$$

where N is the particle density (electrostatic forces will insure approximate equality of ion and electron densities), $\mathbf{V}_d$ is obtained from Eq. (3) applied to ions, and $\mathbf{V}_e$ is obtained from Eq. (3) applied to electrons. Since the centrifugal force on the electrons is small compared with the centrifugal force on the ions, we have

$$\mathbf{j}_\theta \simeq -(NMV_d^2)[(\mathbf{B}\times\mathbf{R})/B^2R^2], \tag{6}$$

where M is the mass of the ions. Since $\mathbf{B}\times\mathbf{R}$ is in the

---

[1] For a review of the United States' program see Amasa Bishop, *The U. S. Program in Controlled Fusion* (Addison-Wesley Publishing Company, Reading, Massachusetts, 1958).

[2] See reference 1, p. 67.

[3] Anderson, Baker, Bratenahl, Furth, Kunkel, and Stone, *Second International Conference on Peaceful Uses of Atomic Energy, Geneva, 1958*, paper No. 373; Anderson, Baker, Bratenahl, Furth, Ise, Kunkel, and Stone, "Study and use of rotating plasma," University of California Radiation Laboratory Rept. UCRL-8067 (December, 1957).

[4] Boyer, Hammel, Longmire, Nagle, Ribe, and Riesenfeld, *Second International Conference on Peaceful Uses of Atomic Energy, Geneva, 1958*, paper No. 2383.

[5] Wilcox, Gow, and Smith, Bull. Am. Phys. Soc. Ser. II, 4, 55 (1959); "The ion magnetron," University of California Radiation Laboratory Rept. UCRL-8579.

[6] M. S. Yoffe (private communication); E. E. Yushmanov, *Plasma Physics and the Problems of Controlled Thermonuclear Reactions*, Vol. IV (to be published in translation by Pergamon Press, New York).

[7] L. Spitzer, *Physics of Fully Ionized Gases* (Interscience Publishers, Inc., New York, 1956).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

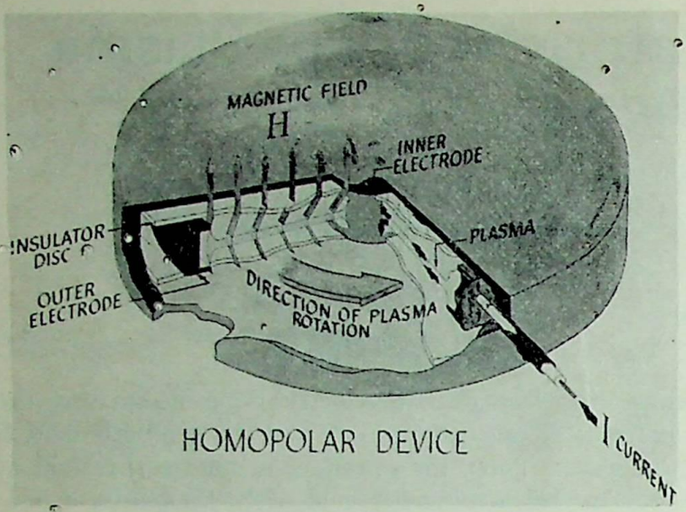


FIG. 1. Diagram of Homopolar device.

same direction as the external solenoidal current that gives rise to **B**, $\mathbf{j}_0$ is a diamagnetic current. For the case mentioned in the foregoing with an outward electric field of 3 kv/cm and a magnetic field of 10 kilogauss, and with a deuteron density of $10^{14}/\text{cm}^3$ and a radius of curvature of 10 cm, the drift-current density is 34 amp/cm^2.

A plasma that is rotating in equilibrium with an applied electric field E can be compared with a condenser that is in equilibrium with an applied electric field; however, in the rotating plasma the energy is stored in the form of kinetic energy of the moving particles. We can derive the dielectric constant k of the plasma by equating this kinetic-energy density to the increased energy density of the electric field,

$$\frac{kE^2}{8\pi}=\frac{E^2}{8\pi}+\frac{1}{2}\rho V_E^2=\frac{E^2}{8\pi}\left(1+\frac{4\pi\rho c^2}{B^2}\right), \quad (7)$$

where ρ is the density of the plasma, and where use has been made of Eq. (2), so that the dielectric constant of the rotating plasma is

$$k=1+(4\pi\rho c^2/B^2). \quad (8)$$

This number can be very large in a practical situation; for the case mentioned in the foregoing the dielectric constant is 3.7×10^4. The effect of the displacement of the magnetic lines has not been included in this derivation.

Post has shown that under certain conditions a charged particle is reflected if it moves into a region of increasing magnetic field, so that particles can be contained if the magnetic field is made to increase at both ends.[8] Post has also shown that in order for a particle to pass through the mirror, we must have

$$W_{11}(c)\geq W_\perp(c)[(B_m/B_c)-1], \quad (9)$$

where $W_\perp$ and W_{11} are the perpendicular and parallel kinetic energies, and c refers to the center and m to the mirror. If the plasma is rotating, the centrifugal force tends to keep the particles away from the axis, where they must go to escape, and thus the mirror containment efficiency is enhanced.[3] An elegant theoretical treatment[4] shows that for a rotating plasma the foregoing inequality is to be replaced by

$$W_{11}(c)\geq W_\perp(c)[(B_m/B_c)-1]+\tfrac{1}{2}MV_E^2(c)[1-(B_c/B_m)], \quad (10)$$

where M is the mass of the particle and V_E is given by Eq. (2). In this case $W_\perp$ is to be measured in the rotating frame of reference. Thus a rotating plasma is more efficiently contained by the magnetic mirror.

Finally, we consider the possible formation of an anode sheath for the case in which the anode surface is parallel to the magnetic field lines. In a conventional electron-discharge tube the sheath forms at the cathode because of the high mobility of the electrons. However, for motion perpendicular to a magnetic field the ions have a greater mobility, while the electrons are tightly confined to their field lines. The ions acquire their electric drift energy by moving a short distance in a direction parallel to the electric field. As they move away from the surface of the anode a region with net negative space charge is left behind. In this narrow sheath a large voltage drop may be present. In some of the experiments the presence of the anode sheath is considered desirable, while in others it is avoided.

THE HOMPOLAR EXPERIMENT

In the Homopolar experiment the emphasis has been on the study of the basic properties of rotating plasmas, therefore argon has been used for the most part. In this experiment[3] the object is to create a spinning disk of plasma which is somewhat similar to the flywheel of a conventional Homopolar generator. The experimental geometry is shown in Fig. 1. The vacuum chamber, which has a diameter of 30 cm and a height of 5 cm, is placed between the pole tips of a magnet so that there is an axial magnetic field of up to 18 000 gauss. A radial electric field exists between the inner electrode and the concentric cylindrical outer electrode. By operation at a comparatively high gas pressure ($\sim$100 μ) the anode sheath effects are largely removed, so that the electric

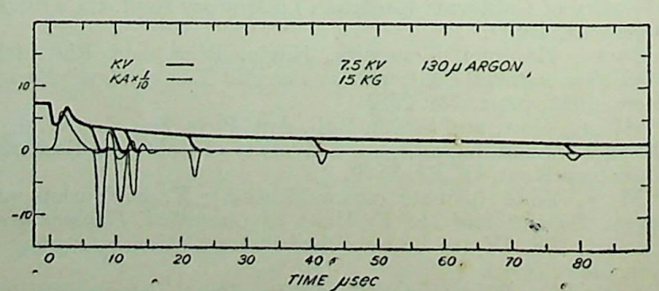


FIG. 2. Current and voltage wave forms in the Homopolar device.

[8] R. F. Post, Bull. Am. Phys. Soc. Ser. II, 3, 196 (1958); *Second International Conference on Peaceful Uses of Atomic Energy, Geneva, 1958*, paper No. 377.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

field exists through the body of the plasma, and the entire plasma is set spinning with the electric drift velocity. Initially a large pulse of current is required to ionize the gas and supply the energy of rotation. Once the plasma has been set in rotation a much smaller current will be drawn, due only to the viscous drag of the plasma. In Fig. 2 the heavy trace shows the behavior of the voltage, while the positive light trace, which extends from 0 to 6 μsec, is the charging-current pulse. The negative current pulses are discussed below.

The equivalent circuit of the Homopolar experiment is shown in Fig. 3. Thus the spinning plasma is represented by a capacitance C_H, which arises from the dielectric effect discussed above, with a parallel resistance R_L, which represents viscous effects. Indeed this geometry may have application as a very-low-inductance capacitor for driving controlled-fusion experiments.[9]

Plasma rotation has been demonstrated by three observations. If we consider the rotating Homopolar as a charged capacitance, then it should be possible to recover the initial charge that was fed in by the pulse of charging current. For this purpose the terminals of the Homopolar are rapidly shorted with the crowbar circuit shown in Fig. 3. The resulting current pulse is in the opposite direction to the charging current, and resembles one of the negative traces shown in Fig. 2. By application of the crowbar at various times, the amount of energy still stored in the system after any interval can be measured. The dissipation of stored energy is indicated by the decreased crowbar current traces at longer times. The crowbar method has shown that under favorable conditions more than half of the input energy can be recovered.

Plasma rotation can also be observed by the Doppler shift of the emitted radiation, as shown in Fig. 4. A spectrometer views the rotating plasma through a tangential port. If the direction of the magnetic field is reversed, then, according to Eq. (1), the direction of rotation will be reversed. The rotational velocities calculated from the Doppler shift agree with the velocities calculated from the crowbar current. Rotational energies of a few hundred electron volts are measured during operation with argon.

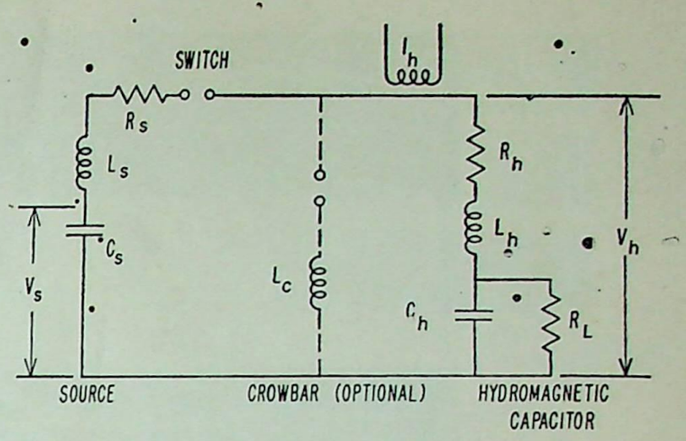


FIG. 3. Equivalent circuit of Homopolar device.

A third observation of plasma rotation is the "spin loop" data. As shown above, the centrifugal force from the plasma rotation gives rise to a diamagnetic current in the θ direction, and thus to a radial component of magnetic field. We may think of the rotating particles as "leaning" on the magnetic lines and bending them outward. This effect is independent of the direction of rotation. Thus the signal in a magnetic probe coil which is oriented to detect H_r should be independent of the sense of rotation. The observed spin-loop signal is shown in Fig. 5, and indeed has the same polarity for both directions of rotation. A slight misalignment of the pickup coil is believed to account for the difference in amplitude between the two signals.

Thus a considerable storage of rotating energy has been demonstrated in the Homopolar experiment in a geometry which is stable for times of the order of 100 μsec. For controlled-fusion applications some of this energy will have to be converted to random or heat energy, which presumably can be accomplished with suitable perturbations in the electric or magnetic fields.

The same basic ideas have been incorporated in the improved geometry shown in Fig. 6. In this case the centrifugal force is more effective in pulling the plasma away from the insulators. Experiments are just beginning with this geometry; however, a decay time of several hundred microseconds has been observed.

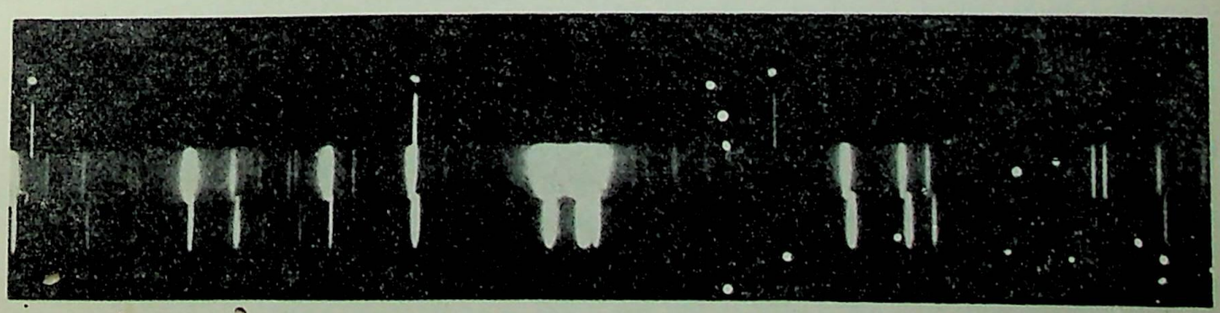


FIG. 4. Doppler-shifted spectra from Homopolar device.

[9] Anderson, Baker, Bratenahl, Furth, and Kunkel, J. Appl. Phys. **30**, 188 (1959).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

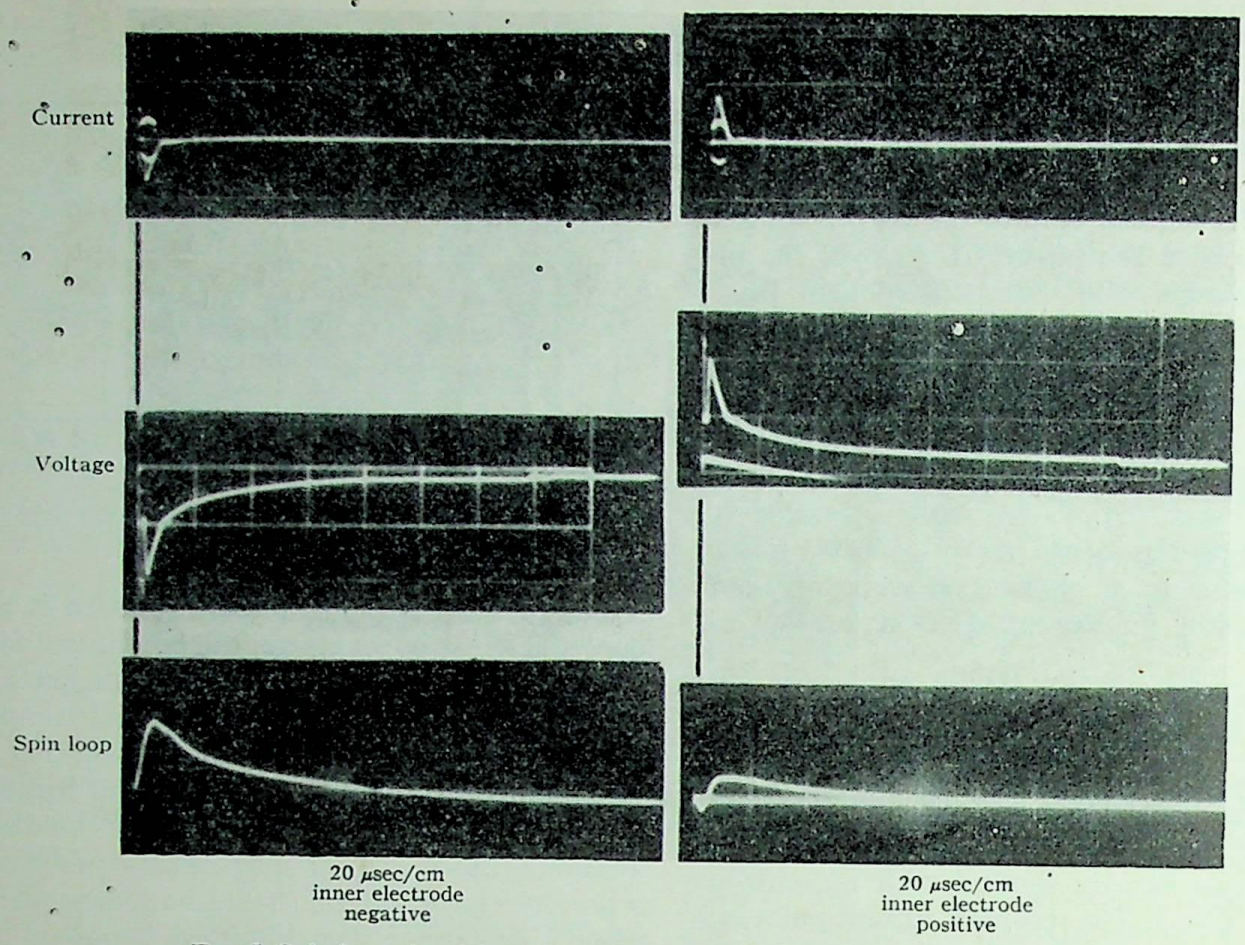


FIG. 5. Spin-loop signal from Homopolar device—100 μ Ar; 17.4 kilogauss; 7 kv.

IXION

This experiment is named for the king in Greek mythology who aspired to the love of a goddess and was therefore sentenced to be forever bound to a flaming rotating wheel.

The Ixion[4] geometry as shown in Fig. 7 is long (86 cm) in comparison with the flat-disk geometry of the Homopolar. The pulsed magnetic field increases from 9 kilogauss at the median plane to 20 kilogauss at the mirror throat, and the pressure between pulses is 2×10^{-6} mm Hg. The first experiments were done with metallic central electrodes, but the latest model has a plasma "center rod." A puff of deuterium gas corresponding to one micron pressure in the apparatus is ionized and driven into the machine by a magnetic shock coil. This plasma electrode makes electrical contact with a pair of tungsten ring electrodes located at the mirror throats. After the plasma electrode has arrived, the high voltage (10 kv) is pulsed on. For about 200 μsec negligible current is drawn, and then a sudden impulse of current of about 20 000 amp is drawn for 25 μsec, as shown in Figs. 8(A) and 8(B), which have a sweep speed of 200 μsec/cm. The voltage falls to about one-half its original value as the discharge occurs, and then decays to zero in about 500 μsec. This current spike and the voltage behavior are interpreted as being caused by the development of a rotating plasma. The subsequent decay of the voltage is attributed to the loss of plasma to the walls.

Plasma rotation has been demonstrated by the crowbar technique and by Doppler-shift observations. When Ixion is short-circuited, about one-fifth of the original charge can be recovered. In the Homopolar experiment one-half the original charge was recovered, but this was

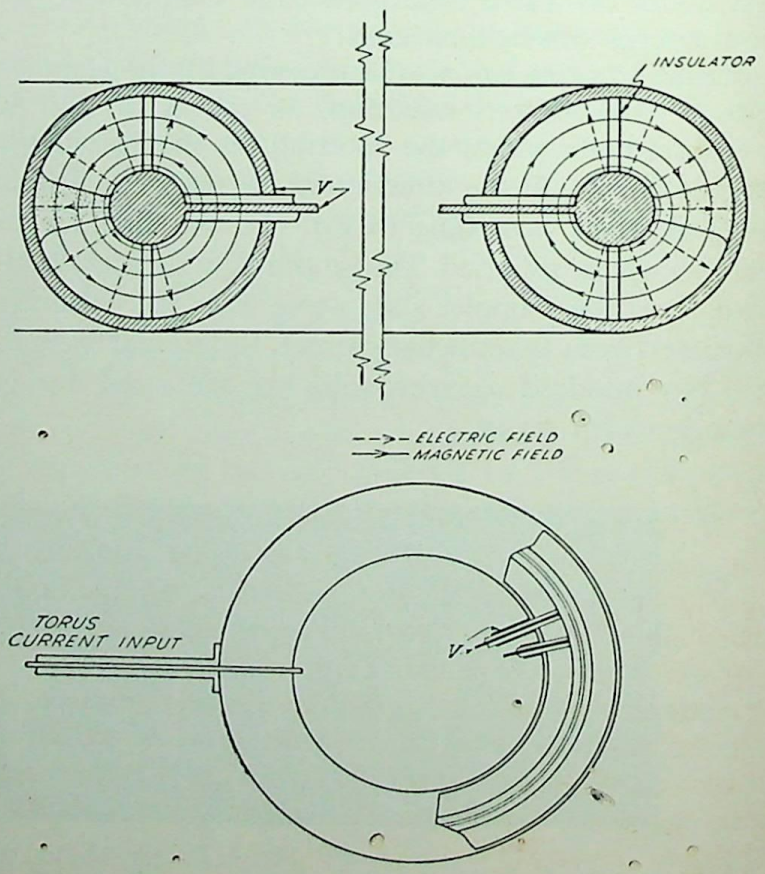


FIG. 6. Later version of Homopolar device.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

with operation with argon, and these machines are found experimentally to operate better with heavy gases than with deuterium. The capacitance of Ixion can be calculated from the plasma dielectric constant [given in Eq. (8)] increased by a factor of the order of 1.5, which corrects for the diamagnetic weakening of the applied magnetic field. Using this capacitance and the observed amount of charge that can be recovered, one can compute that the rotating deuterium plasma would have an equivalent pressure of 5.6 μ. The pressure of the injected deuterium is 1.5 μ; the discrepancy may be accounted for by the presence of impurities or by an error in the geometrical correction.

The Doppler shift that occurs when the magnetic field is reversed is shown in Fig. 8(D), which is a profile of a carbon line. The drift velocity of 4×10^6 cm/sec calculated herefrom is consistent with the value calculated from the applied electric and magnetic fields. The spectral analysis also shows many other impurity lines during the current pulse, and H and D lines lasting for 100 μsec after the current pulse.

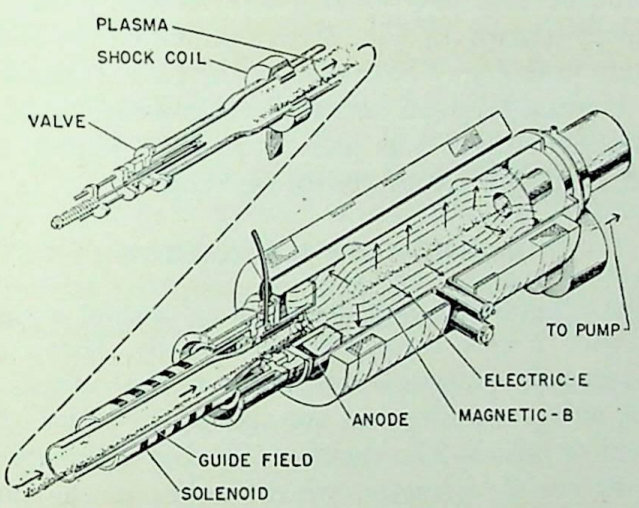


FIG. 7. Diagram of Ixion.

The high-impedance delay for 200 μsec after the voltage is pulsed on was examined with an earlier model having a metallic central electrode. By electrostatic probing it was found that most of the voltage drop occurred just outside the anode, corresponding to the presence of an electron sheath as discussed above. Thus it is seen that the electric field does not penetrate the main body of the plasma and bulk rotation does not occur.

If the magnetic field is lowered, the voltage raised, or the pressure raised excessively, the discharge is oscillatory as shown in Fig. 8(C). The voltage falls to zero almost at once. Neutrons occur during the current spike in this mode of operation, with about 10^5 neutrons per discharge at 40 kv and a detectable yield at 10 kv. If the magnetic field is increased so as to obtain the capacitive-type operation shown in Figs. 8(A) and 8(B), the neutrons disappear. The origin of these neutrons has not been satisfactorily explained.

In an Ixion with a metal center rod, large diamagnetic

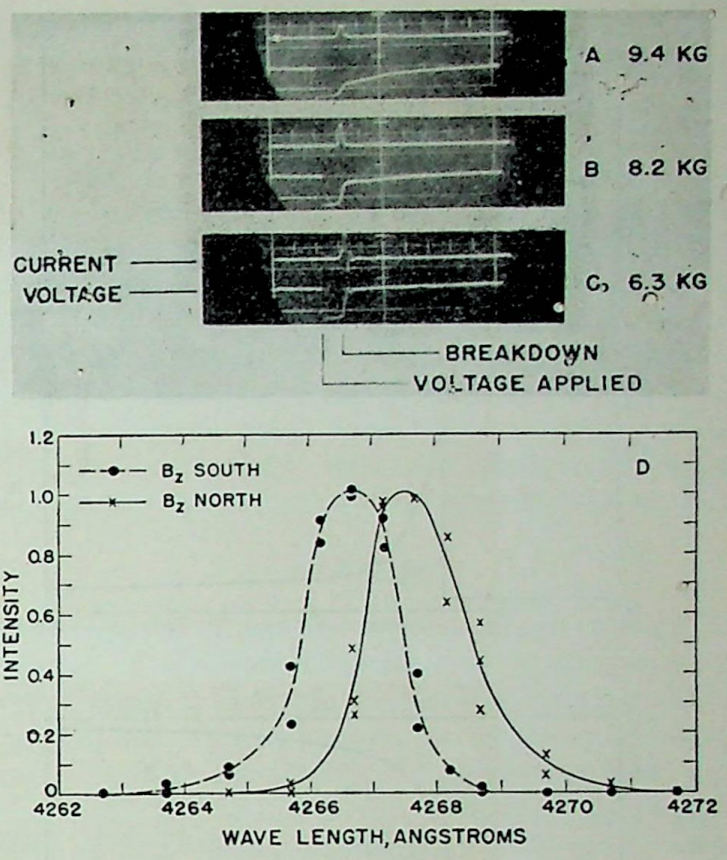


FIG. 8. Experimental data from Ixion.

signals were observed with a magnetic probe coil inside the center rod during the oscillatory discharge of Fig. 8(C). These signals corresponded to almost complete removal of the axial magnetic field from the center of the machine. The diamagnetic effect decreased to zero as the probe was moved away from the center plane to the mirror throat. Diamagnetic signals have also been observed on Ixion with the plasma center rod. The diamagnetic decrease for voltage-holding operation is of the order of 30% at the center, and decreases to zero at the mirror throat.

BERKELEY ION MAGNETRON

This experiment[5] utilizes a long solenoidal magnetic field (120 cm) having a mirror throat at each end and a metal center electrode, as shown in Fig. 9. However, the philosophy on which this is based is quite different from that in the two preceding experiments. The tube is operated at the comparatively low pressure of a

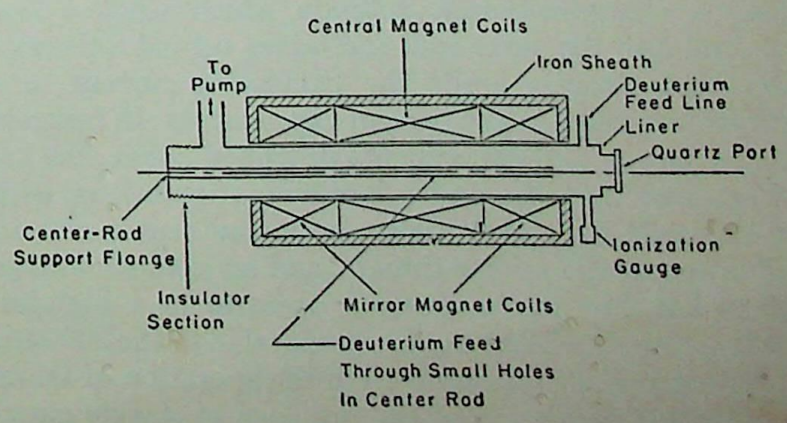


FIG. 9. Diagram of Berkeley ion magnetron.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

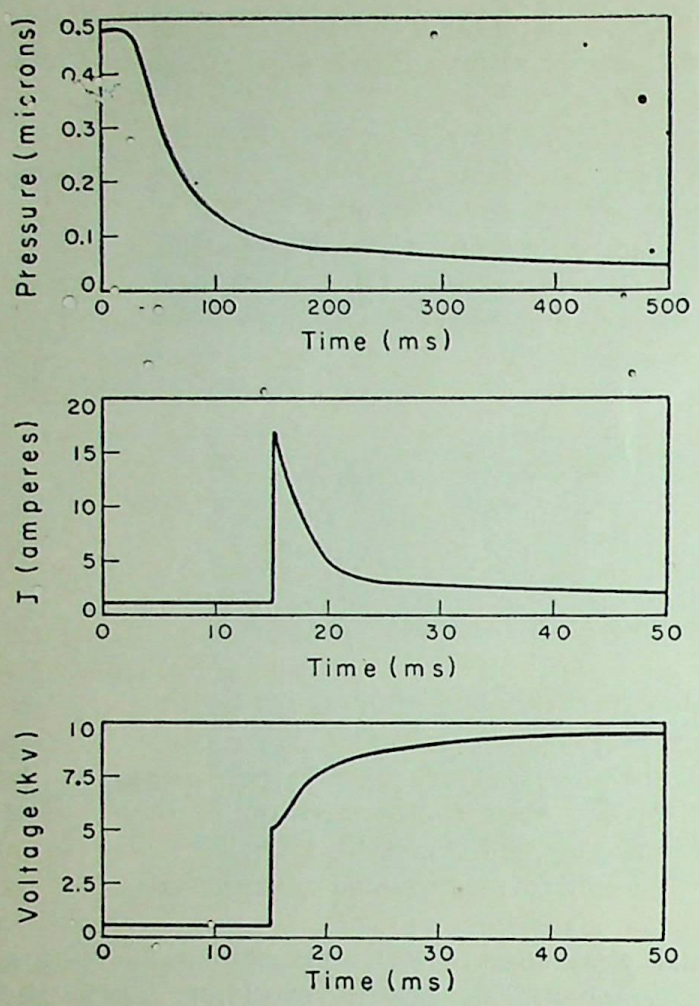


FIG. 10. Experimental data from Berkeley ion magnetron—0.35 μ deuterium; 9 kilogauss; 10 kv.

fraction of a micron, and thus the anode sheath develops about the central electrode. Since this sheath contains a considerable density of ionizing electrons circulating around the axis, a neutral deuterium molecule that enters the sheath volume has a good chance of being ionized. The ion thus formed will be rapidly accelerated out of the sheath by the electric field, and will circulate around the axis in magnetronlike orbits with a kinetic energy equal to the potential energy at the point in the sheath at which it was ionized. The magnetic field is considerably stronger than the magnetron cutoff value, so that the circulating ions cannot reach the outer wall. Each time the ion returns toward the axis it is reflected by the sheath potential; the sheath protects the central electrode. After several revolutions this fast ion undergoes a charge-exchange collision which sends a fast neutral into the outer wall and leaves behind a thermal ion. This process limits the circulating current to a rather low value in the present experiments. In contrast to the two pulsed experiments described above, the ion magnetron operates under continuous conditions, with a magnetic field of 3 kilogauss at the center and 12 kilogauss at the mirror throats, and an applied voltage of 10 kv. The ion magnetron operates in a high-impedance mode which may be comparable to the 200-μsec "waiting period" encountered in the operation of Ixion.

The first indication of the presence of sizable circulating ion currents came with the observation of rather violent oscillations of the center electrode. This electrode was mechanically supported at only one end, so that it was free to oscillate as a pendulum. This motion could be viewed through the quartz port shown in Fig. 9. Several exploratory experiments supported the view that the force responsible for these oscillations came from the recoil of ions accelerated in the sheath. From the magnitude of the forces involved, the circulating current was estimated to be equivalent to about 100 amp when the current drawn by the tube was about 1 amp.

When the tube voltage is turned on, the behavior of the gauge pressure and tube current is as shown in Fig. 10. The size and duration of the initial current pulse can be computed by considering that the neutral gas initially present flows into the sheath with thermal velocity.

With steady operation at a tube voltage of 10 kv and a tube current of 1 amp, neutrons are produced at a rate of a few times 10^4 per sec. They are probably produced by head-on collisions between ions entering the sheath and ions leaving the sheath. When the geometry shown in Fig. 9 was modified to include vacuum pumping at both ends of the tube, the neutron rate increased by an order of magnitude, since the circulating current was able to reach a higher value before being limited by charge exchange.

MOSCOW ION MAGNETRON

This experiment[6] also utilizes a solenoidal magnetic field with mirror throats on both ends. However, a direct-current plasma beam serves as the central electrode, which is positive at voltages up to 40 kv, direct current or pulsed. The vacuum chamber is 170 cm long and 50 cm in diameter, which makes it the largest among the experiments discussed here. The system geometry shown in Fig. 11 has been drawn from a sketch made by Dr. M. S. Yoffe. The central magnetic field is 8 kilogauss, rising to 12 kilogauss at the mirror throats. Oil diffusion pumps at the ends and evaporated-titanium pumps in the middle give a large vacuum-pumping rate and a base pressure of 10^{-7} mm Hg.

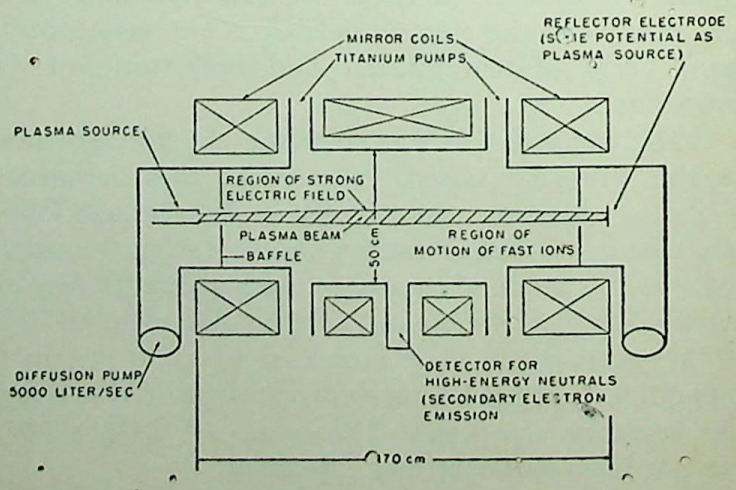


FIG. 11. Diagram of Moscow ion magnetron (Yoffe).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

High-energy neutrals from the charge-exchange process previously mentioned are detected by secondary-electron emission, as shown in Fig. 11. Neutron production is observed with a plastic phosphor viewed by a phototube. In one type of experiment the high voltage is pulsed on with a square wave for from 10 μsec to 2 msec. At the end of the square wave the voltage is returned to ground in about 10^{-7} sec. The high-energy neutral-detector response decreases approximately exponentially with a lifetime of 1 or 2 msec, and neutron pulses are seen for a few msec after the voltage has gone to zero. Thus the decay time of the system appears to be 1 or 2 msec, somewhat longer than has been observed with the Homopolar or Ixion; however, the system dimensions are larger in this experiment, and the mode of operation is considerably different.

This ion magnetron makes 10^7 to 10^8 neutrons per second at 40 kv, with a tube current of 1 or 2 amp. The curve for neutron yield *vs* voltage follows the general shape of the D^+ cross section.

In an earlier smaller model, extensive measurements were made with hot and cold probes of the radial distribution of potential. Three modes of operation may be distinguished. If the applied voltage is very high, all the ions supplied by the plasma source are accelerated radially as soon as they leave the mouth of the plasma source, so that there is no central beam. As the applied voltage is reduced, a second mode of operation is reached in which the central beam reaches a considerable distance into the machine, but not all the way across. In this case a stable radial potential distribution is observed, and most of the electric field is concentrated in the vicinity of the axis, corresponding to the formation of the anode sheath discussed above. Finally, if the applied voltage is decreased still more, a mode of operation is attained in which the central plasma column reaches across the entire length of the machine. In this regime the radial potential distribution oscillates at frequencies of 200 to 400 kc between a steep distribution which corresponds to the presence of an anode sheath and a sloping distribution which corresponds to the absence of a sheath voltage drop. It is not known at present whether these oscillations can be avoided. Since a plasma central electrode offers many advantages, the question of its stability is an important one.

SUMMARY

By emphasis of various aspects of the fundamental concepts of electric drift velocity, diamagnetic azimuthal current, dielectric constant of rotating plasma, magnetic mirror enhancement, and anode sheath formation, the four experimental approaches have been developed. It is clear that this work has only begun to define the problems and possibilities of rotating plasmas. We hope that this article has pointed out some of the challenging experimental and theoretical problems that remain to be solved.

ACKNOWLEDGMENT

We wish to thank Lloyd Smith for a critical reading of the manuscript.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

Ultrasonic Absorption Due to Chemical Relaxation in Electrolytes*

G. S. VERMA

Department of Physics, University of Allahabad, Allahabad, India

TABLE OF CONTENTS

I. Introduction
II. Theory
  (a) Excess Sound Absorption
  (b) Activation Energy for the Rate Process
  (c) Different Mechanisms
III. Methods for Measuring the Absorption
  (a) Methods Based upon Progressive Wave Technique
    (1) Mechanical Methods
    (2) Optical Method
    (3) Electrical Methods
  (b) Methods Based Upon Standing Wave Technique
    (1) Resonance Method
    (2) Reverberation Method
IV. Results and Discussions
  (a) Values of Absorption and Relaxation Frequencies
  (b) Effect of Concentration on Absorption
  (c) Effect of Concentration on Relaxation Frequency
  (d) Effect of Dielectric Constant
  (e) Effect of Temperature
  (f) Effect of Heavy Water When Used as Solvent
  (g) Effect of Pressure

I. INTRODUCTION

STUDY of ultrasonic absorption due to chemical relaxation in electrolytes has recently had considerable development.† As a result of the comprehensive work done by Professor Tamm and his associates at the University of Göttingen and Professor Leonard and his associates at the University of California in Los Angeles, this aspect of sound absorption is understood somewhat better compared to the thermal or structural relaxation in fluids. Recently Litovitz and his associates have contributed much to the understanding of the basic mechanisms responsible for the sound absorption in chemical solutions.

Einstein[1] was the first to point out that chemical processes can account for sound absorption in partially dissociated gases. He made a theoretical estimate of its contribution in nitrogen tetroxide.[2] According to him, the excess sound absorption in partially dissociated gases such as nitrogen tetroxide is due to dissociation.

Liebermann[3] in 1949 considered sound absorption in a chemically active fluid. According to him, chemically active media may exhibit unusual physical properties such as dynamic compressibility whose value depends upon the compression rate in contrast to the conventional static value. The case is similar for dynamic specific heat, which depends upon the rate of change of temperature. These distinctions between static and dynamic values become prominent when high-frequency sound waves pass through the medium which is in chemical equilibrium, provided the equilibrium is pressure and temperature dependent. Such chemical reactions occurring in liquids are expected to affect the propagation of sound waves too.

Acetic acid is the first liquid in which the chemical process was found to be responsible for excess sound absorption. Spakowski[4] and Richards[2] suggested that the absorption of ultrasonic waves arises from the perturbation of the equilibrium between the dimer and monomer forms of acetic acid. Pinkerton[5] in 1948 confirmed experimentally the existence of the relaxation phenomenon in acetic acid and, with Lamb,[6] studied the absorption and dispersion of ultrasonic waves in acetic acid in great detail. They showed that a dispersion and maximum in the value of absorption per wavelength occur and that a relaxation process which arises from the perturbation of the chemical equilibrium due to passage of sound waves does exist. Recent measurements of Piercy and Lamb[7] at low concentrations of acetic acid indicate the presence of a second relaxation region with a characteristic frequency an order of magnitude higher than that for the pure acid. This second relaxation is ascribed to the monomer-dimer reaction, whereas the first relaxation in pure acid is still unexplained. Litovitz and Carnevale[8] have studied effects of pressure on ultrasonic relaxation in pure acid and their results, however, are in agreement with monomer-dimer theory.

* Part of this work was carried out while the author was at Ultrasonics Research Laboratory, Western Reserve University, Cleveland, Ohio.

† Additional references, other than those cited in the text of this article, are listed at the end of the article on p. 1068.

[1] A. Einstein, Preuss. Akad. Wiss. Berlin Ber. **19**, 380 (1920).

[2] W. T. Richards, Revs. Modern Phys. **11**, 36 (1939).

[3] L. Liebermann, Phys. Rev. **76**, 1520 (1949).

[4] B. Spakowski, Compt. rend. acad. sci. U.R.S.S. **18**, 169 (1938).

[5] J. M. M. Pinkerton, Nature **162**, 106 (1948).

[6] J. Lamb and J. M. M. Pinkerton, Proc. Roy. Soc. (London) **A199**, 114 (1949).

[7] J. E. Piercy and J. Lamb, Trans. Faraday Soc. **52**, 930 (1956).

[8] T. A. Litovitz and E. H. Carnevale, J. Acoust. Soc. Am. **30**, 134 (1958).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

Another indication of the existence of such perturbation effects of sound waves on chemical equilibrium resulting in sound absorption is provided by measurements of the ultrasonic absorption in sea water and its comparison with fresh water below one megacycle per second by several workers.[9–14] Leonard discovered that the cause of excess sound absorption in sea water is the presence of $MgSO_4$. Above 1 Mc the absorption is similar to that of fresh water. Below 1 Mc the value of α/ν^2, where α is the absorption coefficient and ν the frequency, is found to increase with decrease in frequency. The absorption coefficient is governed by an empirical relation of the type

$$2\alpha = A_1[K\omega^2/(\omega^2+K^2)]+A_2\omega^2,$$

where ω is the angular frequency and K's and A's are constants. The first term has the form characteristic of a relaxation phenomenon, and this led to the suggestion that chemical relaxation is the process responsible for the excess sound absorption. Wilson[15] later reported that the relaxation frequency does exist near 130 kc. According to Liebermann,[3] the relaxation term arises from a pressure-dependent equilibrium of a chemical reaction within the liquid, and the cause of the excess sound absorption in $MgSO_4$ solutions lies in the dissociation reaction $AB \rightleftarrows A+B$ involving magnesium-sulfate molecules or ions. He developed a theory of sound propagation in chemically active media, but its application to $MgSO_4$ suffered from some errors with regard to concentration dependence; for example, he suggested that the absorption is proportional to the square root of the concentration. Wilson and Tamm and Kurtze[16a] have shown independently that absorption is linearly proportional to concentration. This led to the explanation that the process is unimolecular, $A \rightleftarrows B$, with a relaxing transition between two energy states of the ion pairs. Barthel[17] later discussed the concentration behavior and concluded that if the activity coefficients for ionic reactants are taken into consideration the experimental results of Wilson which suggest unimolecular reaction will fit equally well with the dissociation reaction. Bies[18] also considered the dependence of the activities of the ions with the passage of sound waves. He studied the effect of the dielectric constant of the solvent on the absorption and concluded in favor of a dissociation mechanism as responsible for the excess sound absorption.

Tamm and Kurtze[19] assume the existence of the complex MeR, H_2O, where Me = metal and R = radical. For this complex two different possibilities of dissociation are assumed in order to explain the presence of two relaxation frequencies. The other possibilities which may cause the anomalous sound absorption, according to Tamm, are ion association or a complicated combination of molecular dissociation and hydrolysis. Thermal relaxation is ruled out, simply because there is no appreciable change in the absorption near 4°C where isothermal propagation is expected to take place. The assumption of ion association, i.e., the association of anion and cation to form an ion pair (ion pair $\rightleftarrows$ anion and cation) gives a similar dependence of absorption and relaxation on frequency as in the case of the dissociation process. But this assumption explains only one relaxation mechanism. It does not explain the second increase in the product of absorption cross section and wavelength, which is observed in the case of 2–2-valent electrolytes.

Litovitz and Smithson[20] suggest that all observed data can be explained on the basis of the assumption that the reaction is between two dipole molecules. They rule out the reaction between two ions of unlike sign as the rate-determining step in the primary relaxation process. Their results also indicate that the hydrogen ion is not involved in the rate-determining reaction causing the observed relaxational effects.

II. THEORY

(a) Excess Sound Absorption

A plane acoustic wave in an absorbing medium can be represented in the form

$$\xi_x = \xi_0 \exp[i\omega(t-x/V_0)-\alpha x], \qquad (1)$$

where ξ_x is the amplitude at a distance x and α is the amplitude absorption coefficient. This expression can be written in the form

$$\xi_x = \xi_0 \exp[i\omega(t-x/V)], \qquad (2)$$

where

$$1/V = (1/V_0)-(i\alpha/\omega). \qquad (3)$$

If it is assumed that the chemical equilibrium is pressure dependent, then a change in the pressure will cause a shift in the equilibrium. Such a medium has a complex compressibility. The isothermal compressibility of a solution may be assumed to be equal to the isothermal compressibility of the solvent (assumed real) plus a complex factor due to the perturbed chemical equilibrium. Thus

$$\beta = \beta_0+\beta'. \qquad (4)$$

[9] F. A. Everest and O. H. O Neill, J. Acoust. Soc. Am. **19**, 255 (1946).
[10] D. A. Wilson and L. N. Liebermann, J. Acoust. Soc. Am. **19**, 286 (1947).
[11] R. W. Leonard, J. Acoust. Soc. Am. **20**, 254 (1948).
[12] L. Liebermann, J. Acoust. Soc. Am. **20**, 868 (1948).
[13] Leonard, Combs, and Skidmore, J. Acoust. Soc. Am. **21**, 63 (1949).
[14] R. W. Leonard, Tech. Rept. No. 1, Physics Department, University of California, Los Angeles, California (1950).
[15] O. B. Wilson, Jr., Tech. Rept. No. 4, Physics Department, University of California, Los Angeles, California (1951).
[16] (a) K. Tamm and G. Kurtze, Acustica **3**, 33 (1953); (b) *ibid.*, p. 42; (c) *ibid.*, p. 41.
[17] R. Barthel, J. Acoust. Soc. Am. **24**, 313 (1952).
[18] D. A. Bies, Tech. Rept. No. 6, Physics Department, University of California, Los Angeles, California (1953).
[19] K. Tamm and G. Kurtze, Acustica **4**, 380 (1954).
[20] S. R. Smithson and T. A. Litovitz, J. Acoust. Soc. Am. **28**, 462 (1956).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

The contribution to the intensity attenuation coefficient per wavelength of the complex compressibility is given by

$$2\alpha\lambda = -2\pi \operatorname{Im}(\beta'/\beta_0). \tag{5}$$

Consider the case of a two-state equilibrium reaction

$$A \underset{k_2}{\overset{k_1}{\rightleftarrows}} B.$$

If N represents the molal fraction in the excited state, then at equilibrium

$$dN/dt = k_1 - (k_1+k_2)N = 0. \tag{6}$$

Similarly in the case of a dissociation reaction

$$AB \underset{k_2}{\overset{k_1}{\rightleftarrows}} A+B$$

at equilibrium,

$$dN/dt = k_1(1-N) - k_2N^2. \tag{7}$$

If the equilibrium of the reaction is perturbed only slightly by the change of acoustic pressure, the rate of return towards equilibrium is, to a first-order approximation,

$$d(\Delta N)/dt = K(\Delta N - \Delta N_0), \tag{8}$$

where ΔN_0 represents the magnitude of small perturbation and K is defined as the equilibrium reaction rate. In the case where ΔN_0 is a constant the solution is given by

$$\Delta N = \Delta N_0[1-\exp(-Kt)], \tag{9}$$

where $1/K$ is called the time constant or relaxation time τ of the reaction in the vicinity of equilibrium. However, if the perturbing force is the acoustic wave, the perturbation varies periodically with time, for example, as $\exp i\omega t$, then the solution is given by

$$\Delta N = \Delta N_0/(1+i\omega/K). \tag{10}$$

Liebermann[3] showed that the dynamic partial isothermal compressibility β' is given by

$$\beta' = \beta_0'/(1+i\omega/K), \tag{11}$$

where β_0' is the portion of the static isothermal compressibility which arises from the supposed chemical reaction.

Neglecting the small specific heat term, the contribution to the sound absorption coefficient arising from the chemical reaction is given by

$$2\alpha = \frac{\beta_0'}{\beta_0} \frac{\omega^2 K}{(\omega^2+K^2)C_0}, \tag{12}$$

where β_0 is the total static isothermal compressibility, K is the rate constant for return to equilibrium after a slight perturbation, C_0 is the phase velocity of sound at very low frequencies, and ω is the angular frequency.

Leonard's measurements indicated $\beta'/\beta_0 = 0.5\times10^{-4}$ and $K = 0.9\times10^{-4}$ sec^{-1} in 0.02 M $MgSO_4$ at 25°C.

Liebermann considered two possibilities of the chemical reaction—unimolecular and dissociation type. Corresponding to these types Liebermann derives the following equations for β_0'. For a reaction of the type $A \rightleftarrows B$,

$$\beta_0' = (V/RT)[\Delta V/V]^2 N(1-N), \tag{13}$$

where V is the volume of solution containing 1 M of solute, R is the gas constant, T is the absolute temperature, N is the fraction of molecules in state B at equilibrium, and ΔV is dV/dN. For a reaction of the type $AB \rightleftarrows A+B$,

$$\beta_0' = (V/RT)[\Delta V/V]^2[N(1-N)/(2-N)]. \tag{14}$$

Here N is the fraction of AB dissociated.

Liebermann concluded that anomalous absorption in $MgSO_4$ had its origin in a dissociation reaction and stated that for this reaction K is proportional to m while β_0' would vary as $m^{\frac{1}{2}}$, where m is the molar concentration.

According to Barthel[17] the concentration behavior of β_0' and K is represented as follows. For a reaction of the type,

$$A \underset{k_2}{\overset{k_1}{\rightleftarrows}} B,$$

$$K = k_1+k_2 \tag{15}$$

and

$$\beta_0' \propto k_1k_2(k_1+k_2)^{-2}m; \tag{16}$$

and for

$$AB \underset{k_2}{\overset{k_1}{\rightleftarrows}} A+B,$$

mostly dissociated,

$$K \simeq k_1 \tag{17}$$

$$\beta_0' \propto [(k_2/k_1)\nu_A\nu_B m]m; \tag{18}$$

and for

$$AB \underset{k_2}{\overset{k_1}{\rightleftarrows}} A+B,$$

only slightly dissociated,

$$K \simeq 2(k_1k_2\nu_A\nu_B m)^{\frac{1}{2}} \tag{19}$$

$$\beta_0' \propto [(k_2/k_1)\nu_A\nu_B m]^{-\frac{1}{2}}m, \tag{20}$$

where m represents the molar concentration, k_1 is the rate constant for the forward reaction, k_2 is the rate constant for the reverse reaction, and ν's are activity coefficients. Activity coefficients for ionic reactants generally decrease rapidly with increase of concentration. If $MgSO_4 \rightleftarrows Mg^{++}+SO_4^{--}$ is taken as the mechanism corresponding to the case $AB \rightleftarrows A+B$, mostly dissociated, the data of Harned and Owen[21] show that

[21] H. S. Harned and B. B. Owen, *The Physical Chemistry of Electrolytic Solutions* (Reinhold Publishing Corporation, New York, 1950).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

the mean activity coefficient for the magnesium and sulfate ions varies as $m^{\frac{1}{2}}$ with the result that the concentration dependence approximately cancel out of the factor $\nu_A\nu_B m$ making β_0' proportional to m and K, independent of m even in the dissociation case, similar to unimolecular reaction.

Wilson[15] and Tamm and Kurtze[16a] have applied relaxation theory to chemical equilibrium and obtain the following expressions for the absorption coefficient per wavelength. For the unimolecular reaction

$$A \underset{k_2}{\overset{k_1}{\rightleftarrows}} B,$$

the absorption per wavelength is

$$\alpha\lambda = \frac{\pi(\Delta V)^2}{\beta_0 RT}\frac{CK}{(1+K)^2}\frac{\omega\tau}{1+\omega^2\tau^2}, \tag{21}$$

with the angular frequency of relaxation

$$1/\tau = k_1 + k_2. \tag{22}$$

In these equations C is the total concentration of the solute, ΔV is the difference between the partial molar volumes, $K = C_B/C_A = k_1/k_2$ is the equilibrium constant, k_1 and k_2 are the forward and reverse reaction rates, respectively, β_0 is the static compressibility, ω is the angular frequency, and τ is the relaxation time.

For the dissociation reaction

$$AB \underset{k_2}{\overset{k_1}{\rightleftarrows}} A + B,$$

the results are

$$\alpha\lambda = \frac{\pi(\Delta V)^2}{\beta_0 RT}\frac{C\delta(1-\delta)}{(2-\delta)}\frac{\omega\tau}{1+\omega^2\tau^2} \tag{23}$$

and

$$\begin{aligned} 1/\tau &= k_1 + 2C\delta\Pi k_2 \\ &= k_1\{1+[2(1-\delta)/\delta]\}, \end{aligned} \tag{24}$$

where δ is the fractional dissociation and Π is the mean activity coefficient product.

Bies[18] considers a small unit cube in a homogeneous volume of solution. Supposing that no molecules of solvent nor of solute get in or out of the cube and that the dimensions of the cube, which is in a plane sound field, are small compared to wavelength, then the rate reaction for a dissociation reaction $AB \rightleftarrows A + B$ can be written in terms of the numbers of molecules and ions per unit volume as

$$dn_1/dt = k_1 n - k_2 f_\pm^2 n_1 n_2, \tag{25}$$

where $AB = C(1-\delta) = n/N$, $A = C\delta = n_1/N$, $B = C\delta = n_2/N$, and $f_\pm^2$ = mean squared activity coefficient provided C is the total concentration of A in either the molecular or the ionized state and δ, the fraction of A in ionized state.

At equilibrium

$$dn/dt = -dn_1/dt = -dn_2/dt = 0. \tag{27}$$

Here k_1 and k_2 are assumed to be independent of concentration while $f_\pm^2$ is assumed to be a function of concentration. All the parameters n, n_1, n_2, k_1, k_2, and $f_\pm^2$ are assumed to be pressure dependent and have a harmonic time dependence of the form $\exp i\omega t$ in a harmonic sound field. When sound waves pass through the medium, the concentration of the molecules and the ions varies because of the variation in their numbers and the volume. The contribution of the latter to absorption is, however, negligible.

Thus, the general irreversible rate equation may be represented by a Taylor series in terms of small variations in all of the parameters.

If we consider only linear terms,

$$\begin{aligned} i\omega\Delta n_1 = {} & n\Delta k_1 - f_\pm^2 n_1 n_2 \Delta k_2 + k_1 \Delta n \\ & - k_2 f_\pm^2 (n_1\Delta n_2 + n_2\Delta n_1) - k_2 n_1 n_2 \Delta f_\pm^2 \\ & - (k_1 n - 2k_2 n_1 n_2 f_\pm^2)\Delta v. \end{aligned} \tag{28}$$

By using the mass action equilibrium law

$$K = k_1/k_2 = (f_\pm^2 n_1 n_2)/n, \tag{29}$$

we get

$$n\Delta k_1 - f_\pm^2 n_1 n_2 \Delta k_2 = k_2 f_\pm^2 n_1 n_2 (\Delta K/K). \tag{30}$$

From well-known thermodynamic relations

$$\partial(\Delta F_0)/\partial p = -\Delta V_0 \tag{31}$$

and

$$\Delta F_0 = -RT \ln K, \tag{32}$$

the following expression for ΔV_0 is obtained:

$$\Delta V_0 = (-RT/K)(\partial K/\partial n)(\partial n/\partial p)_T. \tag{33}$$

Here ΔV_0 is considered to be an empirical constant as the associated energy change ΔF_0 represents the free energy change when all the reactants are in their standard state.

The contribution β' to the isothermal compressibility due to a perturbation of the dissociation process is

$$\beta' = -(\partial v/\partial n)(\partial n/\partial p)_T. \tag{34}$$

We define

$$\Delta V = N(\partial v/\partial n)_T, \tag{35}$$

$$\beta' = -(\Delta V/N)(\partial n/\partial p)_T. \tag{36}$$

Hence,

$$\begin{aligned} \Delta V_0 &= (+RT/K)(\Delta K/\Delta n)(N\beta'/\Delta V) \\ \Delta K/K &= +[(\Delta V_0)(\Delta V)]\Delta n/(N\beta' RT) \end{aligned} \tag{37}$$

or

$$= (W\Delta n/N\beta' RT), \tag{38}$$

where $W = (\Delta V_0)(\Delta V)$. Thus

$$n\Delta k_1 - \Delta k_2 f_\pm^2 n_1 n_2 = (nk_1 W\Delta n/N\beta' RT). \tag{39}$$

To estimate the variation in the unit volume with the passage of sound waves we note that the net change in volume Δv must lag behind the acoustic pressure p

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

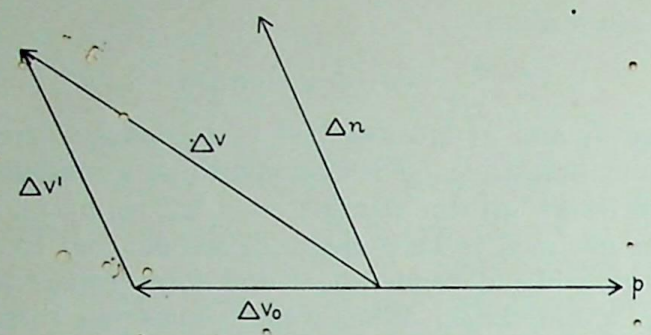


FIG. 1. Vectorial representation of the variations in volume, pressure, and number of molecules n, with the passage of a sound wave.

since this gives the work done on the system as positive. Δv_0 is the variation in volume of the solvent which is assumed dissipationless so that it lags behind the acoustic pressure by 180°. The variation in volume $\Delta v'$ due to a shift in the equilibrium is assumed to be in phase with the variation in the number of molecules Δn. The resultant change in the volume of the solution is Δv. A decrease in volume associated with the dissociation process has been assumed. The foregoing can be represented vectorially in Fig. 1.

From Fig. 1 it follows that

$$\Delta v = -(\beta_0+\beta')\gamma p \tag{40}$$

$$\Delta v' = -\beta'\gamma p. \tag{41}$$

Since the process is assumed to be adiabatic, the ratio of the specific heats γ has been introduced.

Also it may be assumed that

$$\Delta v' = (\partial v/\partial n_2)\Delta n_2 = (\Delta V)\Delta n/N.$$

Eliminating $\Delta v'$ and γp between the last three expressions, we have

$$\Delta v = \Delta V(\beta_0+\beta')\Delta n/N\beta'. \tag{42}$$

To estimate $\Delta f_\pm^2$, consider a function F defined such that

$$F = -(Z_1^2+Z_2^2)d(\ln f_\pm^2)/d\Gamma, \tag{43}$$

where Γ is the ionic concentration and is given by

$$\Gamma = N^{-1}\sum_{i=1}^{i=4} n_i z_i^2. \tag{44}$$

Thus, it can be shown that

$$\Delta f_\pm^2 = -f_\pm^2 F(\Delta n/N - \Gamma'\Delta v), \tag{45}$$

where $\Gamma' = \Gamma/(Z_1^2+Z_2^2)$.

After substituting Eqs. (39), (42), and (45) into Eq. (28) and canceling out Δn, the following expression for β' is obtained:

$$\beta' = k_1 n N^{-1}[W(RT)^{-1} - (\Delta V)\beta_0(F\Gamma'-1)][i\omega - K]^{-1},$$

where

$$K = k_1[1+n(n_1^{-1}+n_2^{-1}) + nN^{-1}(F+\Delta V - F\Gamma'\Delta V)]. \tag{46}$$

If we assume that (ΔV) is very small, such that $(\Delta V)\ll F$ and $(\Delta V)\ll(\Delta V_0)$, all the terms containing (ΔV) may be neglected. Thus,

$$\beta' = [k_1 nW/NRT(i\omega - K)],$$

where

$$K = k_1[1+n(n_1^{-1}+n_2^{-1}-FN^{-1})]. \tag{47}$$

β' has a maximum value at $\omega = K$ and thus K is the relaxation frequency.

The maximum value of $(2\alpha\lambda)$ is given by

$$(2\alpha\lambda)_{\max} = -2\pi_{\max}\,\mathrm{Im}(\beta'/\beta_0),$$

where

$$\beta' = nk_1W/NRT(i\omega - K).$$

Thus, the maximum value is

$$(2\alpha\lambda)_{\max} = (\pi nk_1W/NRTK\beta_0). \tag{48}$$

Expressing in terms of concentration and the degree of dissociation δ, the maximum value of the energy absorption per wavelength is given by

$$(2\alpha\lambda)_{\max} = \mu_{\max} = [\pi C(1-\delta)k_1W/K\beta_0RT] \tag{49}$$

and

$$K = k_1[1+(1-\delta)\delta^{-1}(2-Fc\delta)]. \tag{50}$$

As the function F always appears as $Fc\delta$, it is better to consider the latter. The function F, as expressed earlier, is

$$F = -(Z_1^2+Z_2^2)(d/d\Gamma)(\ln f_\pm^2) = -8(d/d\Gamma)(\ln f_\pm^2),$$

where $Z_1 = -Z_2 = 2$ and

$$\ln f_\pm^2 = -4.606S\Gamma^{\frac{1}{2}}(1+A\Gamma^{\frac{1}{2}})^{-1} \tag{51}$$

according to Debye-Hückel theory. If the ionic diameter is considered equal to 3.08 A, S and A are given by

$$S = 5.132\times10^6(DT)^{-\frac{3}{2}}, \tag{52}$$

$$A = 109.6(DT)^{-\frac{1}{2}}, \tag{53}$$

where D is the dielectric constant, and T the temperature. Thus

$$Fc\delta = 18.424Sc\delta\Gamma^{-\frac{1}{2}}(1+A\Gamma^{\frac{1}{2}})^{-2}. \tag{54}$$

Eigen[22] has considered the general case which is characterized by a consecutive reaction of the form given below:

$$A^{++}+B^{--} \rightleftarrows (AB)_1 \rightleftarrows (AB)_2 \rightleftarrows \cdots \rightleftarrows (AB)_n.$$

According to him the recombination of oppositely charged ions is hindered by hydration layers and can not be treated as a one-step mechanism. It has to proceed stepwise. This system has a spectrum of relaxation times, in which the number of time constants corresponds to the number of independent steps. The relaxation times have to be calculated from the whole

[22] M. Eigen, Discussions Faraday Soc. 17, 194 (1934); 24, 25 (1957); *Chemische Relaxation* (Steinkopf, Darmstadt, 1958); Eigen, Kurtze, and Tamm, Z. Elektrochem. 57, 103 (1953).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

system of linearized rate equations. For example, in the case of a reaction consisting of two steps,

$$\underset{\text{I}}{A^{++}}+B^{--}\underset{k_{21}}{\overset{k_{12}}{\rightleftarrows}}\underset{\text{II}}{AB'}\underset{k_{32}}{\overset{k_{23}}{\rightleftarrows}}\underset{\text{III}}{AB}$$

with $k_{12}, k_{21} \gg k_{23}, k_{32}$, the linearized rate equations are

$$\begin{aligned}\dot{x}_1&=-k_{12}'x_1+k_{21}x_2\\ \dot{x}_2&=k_{12}'x_1-(k_{21}+k_{23})x_2+k_{32}x_3\\ \dot{x}_3&=k_{23}x_2-k_{32}x_3,\end{aligned}\tag{55}$$

where $\delta C_B=\delta C_A=x_1$, $\delta C_{AB'}=x_2$, $\delta C_{AB}=x_3$, and $k_{12}'=k_{12}(\bar{C}_A+\bar{C}_B)$, where the $\bar{C}$'s are the equilibrium concentrations.

The transformation to a new system, $\dot{y}_i=-y_i/\tau_i$, where y_i is a new set of concentration variables instead of the given variables x_i and τ_i, the corresponding relaxation time, may be written in the vectorial forms

$$\mathrm{y}=M\mathrm{x}$$

and

$$\mathrm{x}=M^{-1}\mathrm{y},\tag{56}$$

where the matrices M and M^{-1} are given by

$$M=\begin{vmatrix}1 & 1 & 1\\ \dfrac{1}{1+A} & \dfrac{1}{1+A} & \dfrac{-A}{1+A}\\ -\dfrac{k_{12}'}{k_{12}'+k_{21}} & \dfrac{k_{21}}{k_{12}'+k_{21}} & 0\end{vmatrix}$$

$$M^{-1}=\begin{vmatrix}\dfrac{k_{21}k_{32}}{k_{12}'k_{23}} & \dfrac{1}{1+A} & \dfrac{k_{21}}{k_{12}'+k_{21}}-1\\ \dfrac{k_{32}}{k_{23}} & \dfrac{1}{1+A} & \dfrac{k_{12}'}{k_{12}'+k_{21}}\quad 1\\ \dfrac{1}{1+A} & -1 & 0\end{vmatrix}$$

with

$$A=(k_{32}/k_{23})[1+(k_{21}/k_{12}')].$$

The relaxation times result as eigenvalues of the functional equation

$$\frac{1}{\tau_1}=0,\quad \frac{1}{\tau_2}=\frac{k_{12}'(k_{23}+k_{32})+k_{21}k_{32}}{k_{12}'+k_{21}},\quad \frac{1}{\tau_3}=k_{12}'+k_{21}.\tag{57}$$

The relaxing parts of the compressibility are given by

$$\beta_{0(2)}'=\frac{C_0\beta(1-\beta)}{2-\beta}\,\frac{1+(k_{12}'/k_{21})}{1+(k_{32}/k_{23})}\,\frac{(\Delta V_2)^2}{RT},$$

$$\Delta V_2=\Delta V_{\text{II}-\text{III}}+\Delta V_{\text{I}-\text{II}}\frac{k_{21}}{k_{12}'+k_{21}};$$

and

$$\beta_{0(3)}'=\frac{C_0\beta(1-\gamma)}{2-\gamma}\,\frac{(\Delta V_3)^2}{RT},\quad \Delta V_3=\Delta V_{\text{I}-\text{II}},\tag{58}$$

where $\Delta V_{\text{I}-\text{II}}=V_A+V_B-V_{AB'}$, $\Delta V_{\text{II}-\text{III}}=V_{AB'}-V_{AB}$.

β and γ are degrees of dissociation; $\beta=C_A/C_0$, $(1-\beta)=(C_{AB'}+C_{AB})/C_0$, $\gamma=C_A/(C_0-C_{AB})$, $(1-\gamma)=C_{AB'}/(C_0-C_{AB})$.

The excess absorption per wavelength is given by

$$(\alpha\lambda)=\frac{\pi}{\beta_0}\left[\beta_{0(2)}'\frac{\omega\tau_2}{1+\omega^2\tau_2^2}+\beta_{0(3)}'\frac{\omega\tau_3}{1+\omega^2\tau_3^2}\right].\tag{59}$$

If $k_{12}', k_{21} \gg k_{23}, k_{32}$, the first step of the reaction is identical with the relaxation effect due to τ_3. For $C_{AB}\to 0 (\gamma=\beta=\delta)$ it becomes identical with the binary one-step equilibrium treated by Bies. The other effect (τ_2) corresponds to the final equilibrium between the state AB and the two states A^{++}, B^{--} and AB'. The relaxation time τ_2 is almost independent of concentration, except in a small transition range at concentrations C_A for which $k_{12}'\approx k_{21}$. Such a treatment seems to describe the experimental facts in a more extended concentration range which Bies's theory fails to do. It could not explain the behavior for concentrations above 10^{-2} M, where the relaxation time as well as the absolute value of $(\alpha\lambda/C)$ becomes almost independent of concentration.

(b) Activation Energy for the Rate Process

The temperature dependence of the relaxation frequency may be utilized to determine an experimental activation energy for the rate process.

Consider the reaction $A\to M'\to B$ where M' represents the activated or transition state for the reaction.

According to Eyring's reaction rate theory, the specific rate for the forward reaction is given by

$$k=(k'T/h)(C'/C_A)=(k'T/h)K',\tag{60}$$

where K' is the equilibrium constant for the activation reaction, k' is Boltzmann's constant, h is Planck's constant, T is temperature, and C' and C_A are the concentrations of the indicated species. $k'T/h$ represents the rate at which molecules pass from the activated state. The true equilibrium constant for the activation reaction is

$$K_0'=K'\nu'/\nu_A,\tag{61}$$

where ν' and ν_A are the activity coefficients. Then the reaction rate is given by

$$\begin{aligned}k&=(k'T/h)(\nu_A/\nu')K_0'\\ &=(k'T/h)(\nu_A/\nu')\exp(-\Delta F^{0\prime}/RT),\end{aligned}\tag{62}$$

where $\Delta F^{0\prime}$ is standard free energy of activation.

Hence the specific rates for the forward reaction and

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

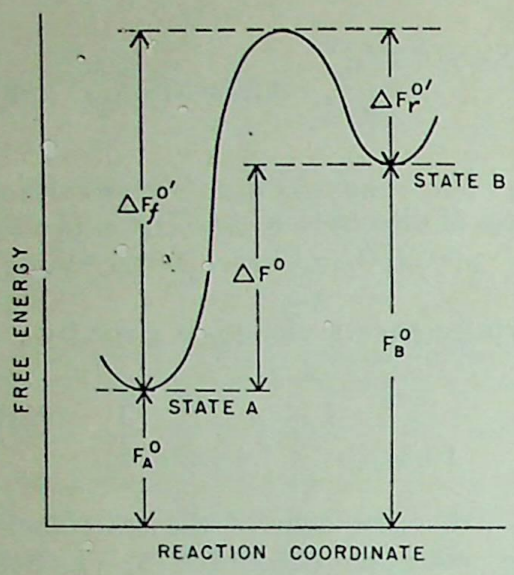


FIG. 2. A two-state potential-energy diagram.

the reverse reaction are

$$k_f = (k'T/h)\exp(-\Delta F_f^{0'}/RT) \quad (63)$$

$$k_r = (k'T/h)\exp(-\Delta F_r^{0'}/RT) \quad (64)$$

considering the activity coefficients as unity. Figure 2 illustrates the relation between the energies in the three states.

The relaxation time τ is given by

$$\tau = 1/(k_f + k_r),$$

where

$$k_f + k_r = (k'T/h)[\exp(-\Delta F_f^{0'}/RT) + \exp(-\Delta F_r^{0'}/RT)].$$

If we assume that $k_f \ll k_r$,

$$\tau = 1/k_r = (h/k'T)\exp(\Delta F_r^{0'}/RT). \quad (65)$$

Hence the relaxation frequency f_m is given by

$$f_m = (k'T/2\pi h)\exp(-\Delta F_r^{0'}/RT). \quad (66)$$

Assuming activation energy independent of the temperature

$$\log(f_m/T) = -(\Delta H_r'/RT)\log e + \text{constant}$$

$$\Delta F_r^{0'} = \Delta H_r' - T\Delta S_r, \quad (67)$$

where $\Delta H_r'$ is the enthalpy and $\Delta S_r'$ is the entropy of activation.

The activation enthalpy can be obtained from the slope of the curve $\log(f_m/T)$ *vs* $1/T$. Since the pdV term is negligible, ΔH is approximately the activation energy.

(c) Different Mechanisms

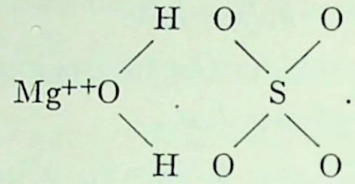
(a) The simplest dissociation reaction for the primary relaxation, which has been suggested by Liebermann,[3] and supported by Leonard, Bies,[18] and Verma and Kor[23] for a 2–2-valent metallic sulfate MSO_4 is

$$MSO_4 \rightleftarrows M^{++} + SO_4^{--}. \quad (68)$$

(b) Barthel[16] has suggested the following mechanism:

$$M^{++} + H_2O \rightleftarrows MOH^+ + H^+. \quad (69)$$

(c) Tamm and Kurtze[19] assume the existence of a complex of the form MeR, H_2O where Me=metal, R=radical. For example, in the case of $MgSO_4$ the complex is represented by



For this complex two different possibilities of dissociation are assumed.

(1) The formation of two univalent ions

$$(MgSO_4, H_2O) \rightarrow MgOH^+ + HSO_4^- \quad (70)$$

which each split up in a subsequent step

$$MgOH^+ + HSO_4^- \rightarrow Mg^{++} + OH^- + H^+ + SO_4^{--}. \quad (71)$$

The first step is considered to be acoustically inefficient. The two dissociation processes are independent of each other and explain the two relaxation frequencies which are observed. For example, in $MgSO_4$ the dissociation of $MgOH^+$ explains the relaxation frequency at 130 kc and the dissociation of HSO_4^- the second relaxation frequency at 200 Mc.

(2) Another possibility exists in the separation of either of the metal ion or of the radical ion from the complex in the form

$$MgSO_4, H_2O \rightarrow Mg^{++} + H_2O \cdot SO_4^{--}$$
$$\rightarrow Mg \cdot H_2O^{++} + SO_4^{--}. \quad (72)$$

(d) Smithson and Litovitz[20] assume the reaction to be between two dipole molecules, i.e., the reaction between the solute molecules (ion pairs) and the dipole solvent. The great excess of the solvent assures that this reaction is of first order. The reaction might be of the type

$$MnSO_4 + H_2O \rightleftarrows Mn \cdot H_2O^{++} + SO_4^{--}. \quad (73)$$

(e) According to Eigen,[22] the simplest possible mechanism allowing an explanation of all hitherto found results may be represented by the following scheme:

$$\underset{\text{I}}{A^{++} + B^{--}} \rightleftarrows \underset{\text{II}}{AB'} \rightleftarrows \underset{\text{III}}{AB''} \rightleftarrows \underset{\text{IV}}{AB'''}. \quad (74)$$

The first equilibrium (I-II) corresponds to diffusion-controlled steps of interaction of the completely hydrated ions, which is present in all electrolytic solutions, especially in those which do not show any detectable absorption below 300 Mc. A relaxation effect due to this interaction should be expected at frequencies higher than 10^3 Mc. The relaxation effects found in 2–2-valent electrolytes are connected with further consecutive re-

[23] G. S. Verma and G. K. Kor, Proc. Phys. Soc. (London) 72, 81 (1958); J. Chem. Phys. 29, 9 (1958).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

action steps in the above scheme (II-III,III-IV). The physical reality of such discrete steps is provided by the fact that the maxima, at least at lower frequencies, have the shape of monochromatic relaxation curves. However, there exists the possibility of more complicated mechanisms in which all the relaxation steps are not detectable.

(f) Recently Verma[24] has proposed a mechanism which shows that when H_2O is replaced by D_2O as the solvent, the primary relaxation frequency should not change, whereas the rate of process responsible for the secondary relaxation is altered considerably. The mechanism is as follows:

$$MSO_4, nH_2O \rightleftarrows [M, (n-1)H_2O]^{++} + [SO_4, H_2O]^{--} \quad (a)$$

$$M^{++} + 2H_2O \rightleftarrows (MOH)^+ + (H_3O)^+ \quad (b)$$

$$SO_4^{--} + (H_3O)^+ \rightleftarrows (HSO_4)^- + H_2O \quad (c)$$

$$MOH^+ + (HSO_4) \rightleftarrows MSO_4 + H_2O. \quad (d)$$

The process (a) with higher activation energy corresponds to the primary relaxation frequency and the process (d) with the lower activation energy to the secondary relaxation frequency. Processes (b) and (c) are acoustically inefficient.

III. METHODS FOR MEASURING THE ABSORPTION

The methods for measuring sound absorption can be classified into two groups depending upon the technique utilized, namely, (1) progressive wave technique and (2) standing wave technique. In the former, absorption is determined from the decrement in the intensity of the plane progressive sound waves governed by the relation

$$I_x = I_0 e^{-2\alpha x}, \quad (75)$$

where I_0 is the intensity at the source, I_x the intensity at a distance x from the source and α the attenuation constant. This is only applicable to measurements above 1 Mc as the absorption, being proportional to the square of the frequency, decreases to such an extent below 1 Mc that the measurements are not reliable. How low is the absorption in water at frequencies below 1 Mc is evident from the fact that the amplitude of a plane wave is reduced only by 10% in a distance of 1 km. Thus, in order to have accurate and reliable results there should be very large volumes of the liquid available in the laboratory, which is hardly possible on such a large scale. Moreover, in order to have the proper directivity of the sound beam, the transducer should be very large.

These difficulties are considerably reduced if absorption is measured by the standing wave technique. In this the absorption is determined from the decay with time of the energy E of a standing wave between the boundaries of the liquid. The decay is governed by the relation

$$E_t = E_0 e^{-2\delta t}, \quad (76)$$

where $\delta = \alpha c$, c being the velocity of propagation of sound waves. The difficulty here lies in the elimination of the boundary losses which influence the decay constant considerably. The decay constant can either be determined by the resonance method or the reverberation method. In the resonance method the decay constant is determined from the decay time of only a single mode of oscillation of the system consisting of the liquid plus the container, whereas in the reverberation method the system is excited to a large number of normal modes of oscillation.

(a) Methods Based upon Progressive Wave Technique

Methods which are based upon the progressive wave technique can be further classified into three main groups which are (1) mechanical methods, (2) optical methods, and (3) electrical methods.

(1) *Mechanical Methods*

The mechanical method, which is in frequent use, depends upon the measurement of sound radiation pressure. When sound waves meet an obstacle the pressure at the front differs from that at the back. The pressure behind the obstacle is assumed to be the pressure in the medium at rest, i.e., in the absence of the sound beam. According to Lord Rayleigh the difference of the pressure, which is known as radiation pressure, is given by

$$S = \tfrac{1}{2}(\gamma+1)(I/C), \quad (77)$$

where γ is the ratio of specific heats, C the velocity, and I the intensity of sound waves. Thus if the radiation pressure is measured at different distances the absorption coefficient for the intensity is easily known from the variation with distance.

The actual radiation pressure, however, as encountered in an acoustic beam under ordinary experimental conditions is different from Rayleigh's pressure. Mean radiation pressure $\bar{P}$ per unit area of a plane compressional wave, however, is defined by

$$\bar{P} = \bar{p} + \langle \rho u^2 \rangle, \quad (78)$$

where p is the excess pressure of the acoustic waves, ρ the density of the medium, and u the velocity. As a generalization of Eq. (78), the mean radiation pressure upon an obstacle of arbitrary shape in an acoustic field can be written

$$\langle \mathbf{P} \rangle = \bar{p}\mathbf{n} + \langle \rho \mathbf{u}(\mathbf{u}\cdot\mathbf{n}) \rangle, \quad (79)$$

where $\mathbf{u}$ now stands for the vector of the velocity $\mathbf{u}(x,y,z)$ and $\mathbf{n}$ is the vector of the inner normal on the surface element under consideration. According to

[24] G. S. Verma, J. Chem. Phys. 29, 1186 (1958).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

Borgnis,[25] the radiation pressure for a beam of finite width in any fluid and on any plane reflecting surface is given by

$$P=2\langle\langle E\rangle\rangle_{\text{kin}}=\langle\langle E\rangle\rangle_{\text{total}}+(\langle\langle E\rangle\rangle_{\text{kin}}-\langle\langle E\rangle\rangle_{\text{pot}}). \quad (80)$$

At small amplitudes $\langle\langle E\rangle\rangle_{\text{pot}}=\langle\langle E\rangle\rangle_{\text{kin}}$ and, therefore,

$$2\langle\langle E\rangle\rangle_{\text{kin}}=\langle\langle E\rangle\rangle_{\text{total}}=\bar{E}_i(1+\beta^2), \quad (81)$$

where $\bar{E}_i$ is the mean energy density of the incident waves and β the reflection coefficient. Hence

$$\bar{P}=2\bar{E}_{\text{kin}}=\bar{E}_i(1+\beta^2)$$
$$=\tfrac{1}{2}\rho_0\omega^2\xi_0^2(1+\beta^2)=(I/C)(1+\beta^2). \quad (82)$$

The above expression is valid both in liquids and gases, when terms of third and higher order in $k\xi_0=(2\pi/\lambda)\xi_0$ are excluded.

There are various ways to measure the radiation pressure. One way uses a torsion balance, which carries on one side a plate or cone which is struck by sound waves, while on the other side weights can be placed to balance the upthrust upon the plate or cone. Sometimes the displacement of the plate or the bead at the end of a long wire when a sound beam traveling horizontally impinges upon it, is used as a measure of sound radiation pressure. Another convenient way to measure the intensity uses a movable vane forming one plate of the condenser, and variations in the capacitance of the system are noted. W. Rufer,[26] employing the above technique, successfully measured the absorption in a large number of electrolytes over a frequency range of 2.9–8.85 Mc for concentrations of 10% and 1 mole/liter.

Ultrasonic absorption in aqueous solutions of $MnSO_4$ has been studied by the pressure balance method in the frequency range (12–43 Mc) by Smith, Barret, and Beyer[27] with the purpose of examining the behavior of aqueous solutions at higher frequencies.

McNamara and Beyer[28] have suggested a variation of radiation pressure method which possesses an accuracy comparable to that of pulse and optical methods. It consists in modulating the transducer at an audio-frequency and detecting the consequent low-frequency variations in the radiation pressure with a condenser microphone.

According to Markham, Beyer, and Lindsay,[29] the measurements with this method below 3 Mc are not reliable since difficulties arise such as those due to retarding forces which might be comparable to the forces which are being measured, specular reflections from the walls of the container, hydrodynamic flow, failure of the detector to intercept the divergent sound beam at low frequencies, and cavitation at high intensities.

Another convenient and accurate method of measuring the absorption coefficient is provided by the velocity of acoustic streaming in a fluid. It is now a well-established fact that there exists a time-independent gradient of radiation pressure associated with the propagation of plane progressive waves of sound through an attenuating fluid medium. This constant gradient produces an acceleration of the fluid and hence streaming. Viscous forces arise which resist the streaming till the steady state condition is reached when these forces exactly balance the gradient of radiation pressure. According to Eckart[30] (1948), Liebermann[31] (1949), Karim and Rosenhead[32] (1952), and Karim[33] (1953), measurements of the streaming velocity provides an independent method of determining the ratio of the second coefficient of viscosity to coefficient of shear viscosity. On the other hand, Nyborg[34] (1953) and Piercy and Lamb[35] (1954) have established both by theory and experiments that the sound attenuation rather than the second viscosity is the factor responsible for streaming. This conclusion was originally pointed out by Fox and Herzfeld[36] (1950). Markham[37] (1952) too supported this view.

The experimental arrangement for measuring the sound absorption in liquids at frequencies around 1 Mc is shown in Fig. 3. The main tube is completely filled by the beam of progressive sound waves and the side tube provides the return path of the streaming where progressive sound waves do not extend.

According to Poiseuille, the flow velocity v_0 of a fluid along the axis of a tube of radius r and length l is given by

$$v_0=(\Delta p r^2/4\eta l), \quad (83)$$

where Δp is the difference in pressure between the ends of the side tube and η is the coefficient of shear viscosity of the fluid, provided the effect of the corners of the side

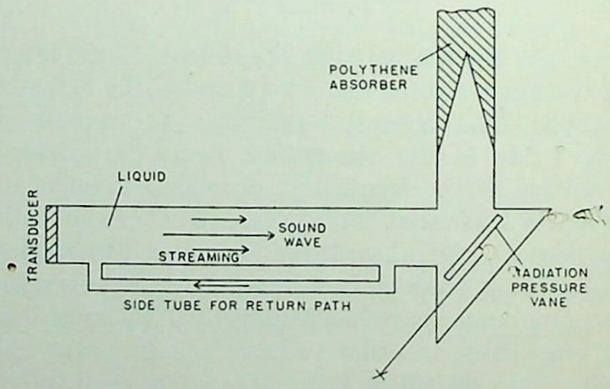


FIG. 3. Sound absorption measurement by acoustic streaming method.

[25] F. E. Borgnis, Revs. Modern Phys. 25, 653 (1953).
[26] W. Rufer, Ann. Physik 41, 301 (1942).
[27] Smith, Barret, and Beyer, J. Acoust. Soc. Am. 23, 71 (1951).
[28] F. L. McNamara, and R. T. Beyer, J. Acoust. Soc. Am. 25, 259 (1953).
[29] Markham, Beyer, and Lindsay, Revs. Modern Phys. 23, 353 (1951).
[30] C. Eckart, Phys. Rev. 73, 68 (1948).
[31] L. Liebermann, Phys. Rev. 75, 1415; 76, 440 (1949).
[32] S. M. Karim and L. Rosenhead, Revs. Modern Phys. 24, 108 (1952).
[33] S. M. Karim, J. Acoust. Soc. Am. 25, 997 (1953).
[34] W. L. Nyborg, J. Acoust. Soc. Am. 25, 68 (1953).
[35] J. E. Piercy and J. Lamb, Proc. Roy. Soc. (London) A226, 43 (1954).
[36] F. E. Fox and K. F. Herzfeld, Phys. Rev. 78, 156 (1950).
[37] J. J. Markham, Phys. Rev. 86, 497 (1952).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

tube on the flow is neglected, which is true for small velocities of streaming. The difference in pressure Δp arises because of the change in radiation pressure due to attenuation in sound waves over a length l in the main tube. Thus, the streaming velocity on the axis of the tube is given by

$$v_0 = \frac{E_0 r^2 (1 - e^{-2\alpha l})}{4\eta l}, \quad (84)$$

where E_0 is the acoustic energy density at the entrance to the side tube nearer the source, which is experimentally known by using the radiation pressure vane. Small aluminum particles are suspended in the liquid, and the velocity of flow is determined by noting their motion through a microscope and η is taken from published tables. According to Piercy, absorptions accurate within 5% can be measured. Recently Hall and Lamb[38] have extended the frequency range of this method down to 100 kc, with an accuracy of ±11% at 130 kc.

(2) *Optical Method*

The optical method is based on the Debye-Sears effect and was first of all developed by Bicquard.[39] Recently many improvements in the general technique have been made by Burton,[40] Willis,[41] and Sette.[42] Kurtze and Tamm used the optical method for the measurement of ultrasonic absorption in electrolytic solutions in the frequency range 3–200 Mc.

The experimental arrangement used by Tamm and Kurtze for the frequency range 3–15 Mc is shown in Fig. 4. Sound beam from the quartz crystal is crossed perpendicularly by a parallel beam of light which, by moving the mirrors can be shifted along it. The liquid tank is 1 m in length and has rubber wedges placed at its ends to avoid the formation of standing waves. The first order of the resulting diffraction pattern is given to a photoelectric cell which is followed by a multiplier. The output of this multiplier is connected to a logarithmic level recorder. The sound is modulated and the multiplier is also tuned to the modulation frequency. In the frequency range 20–100 Mc the principle of the method is the same except that the vessel itself is moved instead of a light beam.

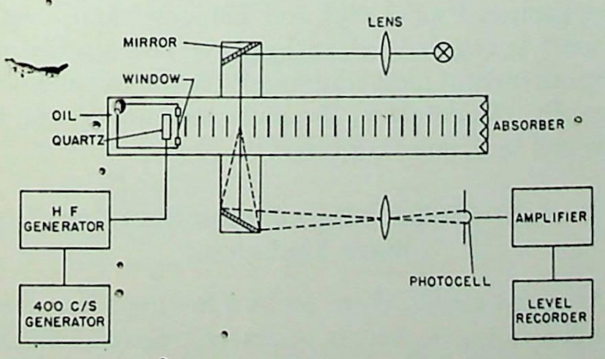


FIG. 4. Sound absorption measurement by the optical method.

(3) *Electrical Methods*

There are three general electrical methods for measurement of ultrasonic absorption in fluids based upon the progressive wave technique. They are (1) interferometric methods, (2) direct methods, and (3) pulse method.

The interferometric methods consist essentially of a plane quartz crystal which acts as a transducer and a reflector set parallel and opposite to the radiating face of the crystal. The distance between them can be varied either by moving the reflector or the transducer itself. Sound waves emanating from the crystal are reflected from the reflector, and stationary waves are formed in the space in between them. If the reflected waves reaching the radiating face are 180° out of phase with the signal just coming out, distrubance just at the radiating face will be reduced to zero and the plate current in the output stage of the driver oscillator will increase. As the reflector is moved towards or away from the source, plate current increases every time the reflector is moved away a half wavelength. According to Pielemeier,[43] the current reading in the output stage of the driving oscillator is proportional to excess pressure in the sound beam at the face of the crystal. Thus, if x represents the distance through which the reflector is moved, the increase in path length will be $2x$ and the pressure of the returning wave is decreased by $e^{-2\alpha x}$. Hence the absorption can be determined by noting the current at two maxima and its decrement with distance.

The whole picture is not so simple, for it has been shown by Hubbard[44] and others that the current is a complicated function of the absorption coefficient, the reflection coefficient of the reflecting surface, and the distance moved by the reflector. Hubbard used a buffer amplifier between the oscillator and the interferometer to reduce the effects of load variations on the frequency of the oscillator. Alleman[45] showed that lack of perfect alignment between the transducer and the reflector leads to erroneous results. Pumper[46] has discussed the departure of the sound beam from a plane wave. Some uncertainty regarding the reflection also remains, as it is alleged to be modified by heat conduction. This method has been used by several workers[47–49] including the author.[50]

In the *direct method* the excess pressure or the in-

[38] D. N. Hall and J. Lamb, Proc. Phys. Soc. (London) **73**, 354 (1959).
[39] P. Bicquard, Compt. rend. **196**, 257 (1933).
[40] C. J. Burton, J. Acoust. Soc. Am. **20**, 186 (1948).
[41] F. H. Willis, J. Acoust. Soc. Am. **19**, 242 (1947).
[42] D. Sette, Nuovo cimento **7**, 55 (1950); J. Chem. Phys. **18**, 1592 (1950).
[43] W. H. Pielemeier, Phys. Rev. **34**, 1184 (1929).
[44] J. C. Hubbard, Phys. Rev. **38**, 1011 (1931); **41**, 523 (1932).
[45] R. S. Alleman, Phys. Rev. **55**, 87 (1939).
[46] E. J. Pumper, J. Phys. U.S.S.R. **1**, 411 (1939).
[47] F. E. Fox, Phys. Rev. **52**, 973 (1937).
[48] J. L. Hunter, J. Acoust. Soc. Am. **13**, 36 (1941).
[49] J. L. Hunter and F. E. Fox, J. Acoust. Soc. Am. **22**, 238 (1950).
[50] G. S. Verma, J. Chem. Phys. **18**, 1352 (1950).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

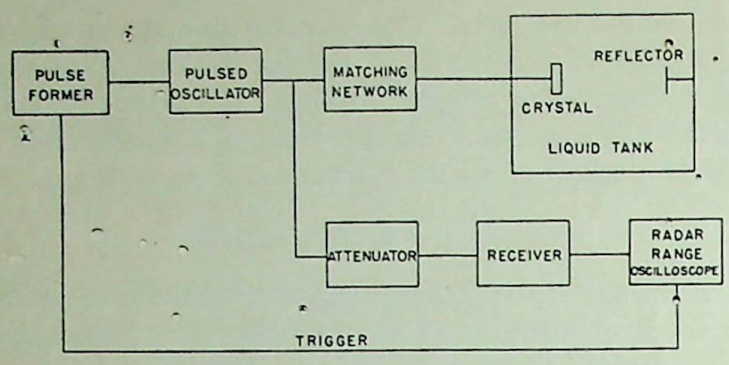


FIG. 5. Sound absorption measurement by the pulse method.

tensity is measured by the microphone which is placed along the axis of the sound beam. The method is not so simple as it appears to be, as many difficulties are associated with it. In the case of the transducer which is assumed to be a pistonlike source, Fraunhofer diffraction takes place within a distance of a^2/λ from the crystal, where a is the radius of the transducer and the typical Fresnel pattern is observed at much greater distances. King[51] and Born[52] have deduced approximate theoretical expressions for the radiation field which can be used for the determination of the absorption coefficient. Correction has to be made for the finite size of the microphone in case its dimensions are not small compared to the wavelength. Standing waves as usual offer a problem which can be reduced by tilting the microphone, so that the face of the microphone is not perpendicular to the axis of the beam.

The direct method has been successfully utilized by Smith, Barrett, and Beyer[27] on the determination of ultrasonic absorption in $MgSO_4$ in the neighborhood of 3 Mc. The detector consisted of a barium titanate crystal connected in parallel with a selenium detector. These two units are mounted in the edge of a probe and connected with a Rubicon galvanometer. The response of the rectifier as read on the galvanometer is proportional to the average intensity of the sound beam over crystal face.

Any method which utilizes continuous waves suffers from two defects. First, there always exists the possibility of stationary waves being formed, which leads to erroneous results of absorption, and secondly, local heating due to sound energy in the main beam may cause temperature waves from the center of the sound beam, which is at a higher temperature, towards the edges of the beam, which is comparatively at a lower temperature. Because of the change in temperature, the temperature recorded by the thermometer kept at a certain distance from the main beam is not the true temperature at which absorption is being measured. But these difficulties are removed in the pulse method.

The experimental arrangement used by Pellam and Galt[53] for the measurement of ultrasonic absorption in liquids is shown in Fig. 5. The arrangement consists of a quartz transducer and polished reflector set up parallel to each other. The pulsed radio-frequency voltage at 15 Mc with a repetition rate of the order of 1000 per sec and pulse width of 1 μsec was applied to the quartz crystal. Thus the average power is only 1/1000 that of a continuous signal at the same amplitude. The same crystal which acts as a transducer is used as a receiver. The distance between the crystal and the reflector can be arranged such that the reflected signal arrives at the instant when there is no transmission of pulse so that stationary waves are not formed. "The scheme is essentially to use the liquid sample as a storage medium for short sound pulses and to measure the time delay and the attenuation undergone by the sound in traversing a known path within the liquid. The acoustical pulses are generated from electrical pulses by means of a crystal transducer and are converted back to electrical form upon completing a transit through the liquid. The effect of changing the acoustic length can be compensated for electrically. Hence, the measured delay produced by the increase in path length gives a measure of the sound velocity; the attenuation which must be removed from the electrical circuit to balance the acoustical losses in additional distances provides a measure of the absorption. With the help of suitable electronic equipment the central and the reflected pulses are compared on an oscilloscope. The initial pulse is generally allowed to pass through a calibrated attenuator and the attenuation required to reduce the size of the initial pulse to match that of the reflected pulse measures the power loss in transmission plus losses due to reflection. When the reflector is moved through a distance x it will mean an increase in path length by $2x$. Thus, if the attenuation noted is plotted against the distance $2x$, the absorption coefficient can be determined from the graph."[53]

Litovitz and his associates have studied ultrasonic absorption in $MnSO_4$ and NH_4OH by the pulse method. Lamb and his associates have utilized pulse technique in studying the relaxation due to rotational isomerism in unsaturated aldehydes and ketones[54] and triethylamine.[55] Young and Petrauskas have also studied rotational isomers of methylbutanes[56] by the pulse method. Recently, the details of the pulse technique have been reviewed by Lamb and others.[57]

(b) Methods Based upon Standing Wave Technique

As stated earlier, there are two methods based upon standing wave technique—namely, the reverberation method and the resonance method. These are frequently

[51] L. V. King, Can. J. Research 11, 135; 11, 484 (1934).
[52] H. Born, Z. Physik. 120, 383 (1943).
[53] J. R. Pellam and J. K. Galt, J. Chem. Phys. 14, 608 (1946).
[54] M. S. de Groot, and J. Lamb, Trans. Faraday Soc. 51, 1676 (1955).
[55] E. L. Heasell and J. Lamb, Proc. Roy. Soc. (London) A237, 233 (1956).
[56] J. M. Young and A. A. Petrauskas, J. Chem. Phys. 25, 943 (1956).
[57] Andreae, Bass, Heasell, and Lamb, Acustica 8, 131 (1958).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

used for the measurement of ultrasonic absorption below 1 Mc. The reverberation technique has been used by Mulders,[58] Moen,[59] and Tamm and Kurtze.[16a] The resonance method has been used by Leonard, Bies, and Tamm and Kurtze.

(1) *Resonance Method*

The *resonance method* differs from the reverberation method in several aspects, the most important of which is the use of a single resonant mode of the spherical resonator. "A small piezo-electric crystal transducer attached to the wall of a liquid-filled spherical resonator is used to excite one of the natural modes of the resonator until the sound intensity reaches a sufficiently high level. The transducer is then switched from the exciting amplifier to a receiving amplifier. The sound energy level now in the resonator decays exponentially at a rate determined by the energy losses in the system. A high-speed power level recorder which plots the logarithm of the decaying receiver output voltage against time, yields a straight decay curve, the slope of which is easily measured. If the losses of the sound energy at the walls and supports of the resonator are negligible, the decrease in sound level results only from the absorption of sound in the contained liquid, and thus the decay rate is a measure of the sound absorption coefficient for the liquid. The decay rate in decibels per second divided by the sound velocity yields the plane wave absorption coefficient in decibels per unit of distance." The schematic diagram of the apparatus is shown in Fig. 6.

The system is normally in the receiving condition as shown. Boundary losses which are an important factor can be eliminated by subtracting the decay constant corresponding to the case when the resonator is filled with a standard liquid. Thus the true absorption of the electrolyte is known from

$$\alpha_{\text{electrolyte}}=\alpha-\alpha_{\text{standard}}=(\delta-\delta_{\text{standard}})/C. \quad (85)$$

The advantages of having a spherical glass resonator are many. First, the deformation and wall frictional losses are minimized by making the walls extremely thin, which is possible in this stable shape. Secondly, there exist certain radial modes of vibrations of which boundary losses can be neglected. Thirdly, it is possible to estimate theoretically the boundary losses considering the deviations of the sound field in the case of an ideal spherical resonator. Bies used round-bottom twelve-liter Pyrex boiling flasks as resonators.

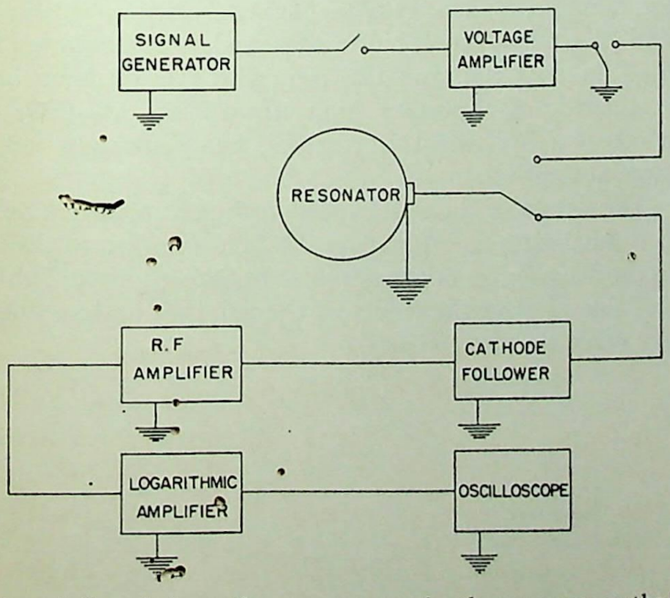


FIG. 6. Sound absorption measurement by the resonance method.

(2) *Reverberation Method*

In the reverberation method the vessel which is filled with the liquid is excited to a large number of normal modes of oscillations so that a diffuse sound field occurs. The absorption coefficient is determined from the rate of decay of the sound. The decay of sound energy is exponential provided the boundary losses affect all the normal modes of vibration nearly uniformly. The nature of the sound field does not matter so long as the sound field is almost uniform throughout the volume of the liquid. If the walls of the container are irregular or very symmetrical, standing wave patterns do not offer any serious problem. Scattering of the sound field, instead of leading to erroneous absorption, helps in achieving the diffuse sound field. The vessel used by Tamm and Kurtze is a seamless cylindrical 100-liter vessel (54 cm in diam and in height) made of pure anodized and varnished aluminum, 3.5 mm thick.

According to Mulders[58] the time constant of the decay of sound amplitude not only depends upon the absorption of ultrasonic waves in the liquids but also on other causes such as the radiation of sound energy into the air and the energy losses by the friction of sound waves along the walls. According to him the decay constant can be expressed as a function of the dimensions of the vessel. In the case of a cylindrical vessel of radius R and height H filled with a liquid of density ρ and shear viscosity coefficient η, the observed decay constant δ is given by

$$\delta=\alpha C+1.18[(2/R)+(1/H)](f\eta/\rho)^{\frac{1}{2}}+(140/\rho H), \quad (86)$$

neglecting the absorption in the attachment of the vessel and the microphone, etc. The second term represents corrections due to the friction of the walls. The third term represents the correction due to radiation into air from the liquid surface, a factor which is not very important. In the case of a cylindrical vessel filled with water, Mulders found at 25°C and 1010 kc $\delta=58.5$ sec^{-1}, $1.18[(2/R)+(1/H)](f\eta/R)^{\frac{1}{2}}=26$ sec^{-1}, $(140/\rho H)=1.5$ using the constants $R=0.14$ m, $H=0.10$ m, $\rho=1000$, $\eta=0.89\times10^{-3}$, $C=1500$ msec^{-1}. Thus αC comes out to be equal to 31.0 sec^{-1} and $(\alpha/\nu^2)\times10^{15}$ m^{-1}sec^2

[58] C. E. Mulders, Appl. Sci. Research **B1**, 149 (1949); **B1**, 341 (1950).

[59] C. J. Moen, J. Acoust. Soc. Am. 23, 62 (1951).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

TABLE I. Values of the absorption coefficient α in decibels per unit length as referred to 0.1-M/l solution at 20°C in the electrolytes investigated so far.

Electrolytes	10 kc/sec db/km	100 kc/sec db/km	1 Mc/sec db/m	10 Mc/sec db/cm	100 Mc/sec db/cm
H_2O	0.0217	2.17	0.217	0.217	21.7
$Al_2(SO_4)_3$	$\geqq$0.1	10.0	1.0	1.0	57
$BeSO_4$	1.75	4.8	0.29	0.29	28
$CaCrO_4$		$\geqq$3.1	0.31	0.31	29
$Ca(CH_3COO)_2$	$\geqq$0.036	3.6	0.36	0.36	36
$CoSO_4$	1.77	162	2.3	0.39	31
$CuSO_4$	0.282	23	0.43	0.39	39
$MgCrO_4$	$\geqq$0.72	54.7	0.48	0.29	29
$MgSO_4$	3.92	247	0.9	0.3	29
MgS_2O_3	1.42	90	0.57	0.28	23.8
$MnSO_4$	0.55	51	5.2	0.74	29
$NiSO_4$	31.0	86	0.48	0.39	29
$ZnSO_4$	$\geqq$0.05	5	0.5	0.5	36
H_2SO_4			0.24	0.24	24
$AlCl_3$			0.223	0.223	22.3
K_2SO_4			0.227	0.227	22.7
K_2CrO_4			0.226	0.226	22.6
$La(NO_3)_3$				0.5	34
Li_2CO_3			0.238	0.238	23.8
Li_2SO_4			0.229	0.229	22.9
Na_2CO_3			0.24	0.24	24
Na_3PO_4		3.9	0.39	0.35	30.7
Na_2SO_4			0.25	0.25	25
$(NH_4)_2SO_4$			0.227	0.227	22.7

$=20.5$ in agreement with measurements of other workers utilizing other methods. Thus, all that is necessary is one measurement of the reverberation time, and other corrections can be calculated from the dimensions of the vessel and the physical constants of the liquid, namely, the viscosity η and the density ρ.

The schematic arrangement of the apparatus is shown in Fig. 7. After the sound source has been stopped, the sound reverberates in the vessel. This can be observed by a crystal microphone and an amplifier. The time constant of decay of sound can be measured at the output of the amplifier. Tamm and Kurtze used the frequency band for excitation of width 20 kc. This is generated by pulsing a variable carrier frequency which is restricted to five definite values, i.e., 50, 100, 200, 500, and 1000 kc. The electric output of a piezoelectric receiver, which is a measure of the sound energy in the vessel (i.e., after amplification, frequency conversion, and filtering) is recorded by a Newman type high-speed level recorder or by a log amplifier combined with an oscilloscope.

Lamb and Sherwood[60] and Karpovich[61] used this technique in the study of relaxations due to rotational isomers in cyclohexane derivatives.

IV. RESULTS AND DISCUSSIONS

(a) Values of Absorption and Relaxation Frequencies

1–1-valent electrolytes, for example, NaCl, NaBr, and KBr, have very low absorption, i.e., the solutions of these electrolytes have almost the same absorption as that of the pure water. 2–1- and 3–1-valent electrolytes as well as 1–2-valent electrolytes behave almost in a similar way to 1–1-valent electrolytes below 10 Mc. Only above 10 Mc do they exhibit measurable absorption.

2–2-valent electrolytes, in general, possess a considerably high absorption in the whole frequency range which has been investigated. In view of the fact that sulfates of bivalent metals are soluble in water and are obtained in sufficient quantities with necessary degree of purity, mostly the sulfates have been investigated; for example, $BeSO_4$, $NiSO_4$, $MgSO_4$, $CoSO_4$, $MnSO_4$, $ZnSO_4$, $CuSO_4$, etc. Thiosulfates and chromates such MgS_2O_3, $MgCrO_4$, and $CaCrO_4$, which have been investigated, also possess high absorption. $Al_2(SO_4)_3$, which is a 3–2-valent electrolyte, has strikingly very large absorption.

The values of the absorption coefficient α in decibels per unit length as referred to 0.1-M/l solution at 20°C in the electrolytes investigated so far are given in Table I. Table II gives the values of the primary and secondary relaxation frequencies.

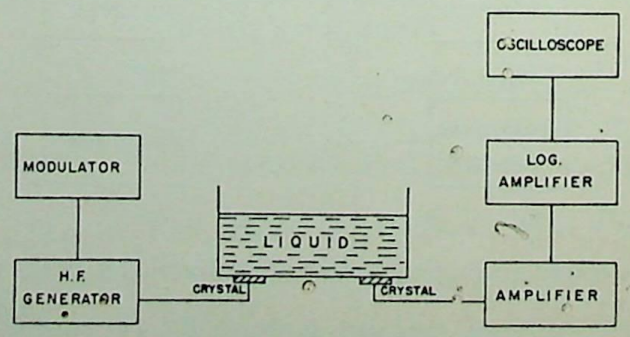


FIG. 7. Sound absorption measurement by the reverberation method.

[60] J. Lamb and J. Sherwood, Trans. Faraday Soc. **51**, 1674 (1955).
[61] J. Karpovich, J. Chem. Phys. **22**, 1767 (1954); J. Acoust. Soc. Am. **26**, 819 (1954).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

Figure 8(a) represents a summary of all measured frequency curves by Tamm and Kurtze at 0.1 M/l. Figure 8(b) shows α/ν^2 for 0.1-M/l solutions plotted against frequency (Tamm and Kurtze[16b]).

(b) Effect of Concentration on Absorption

In nearly all of the 2–2-valent electrolytes investigated the absorption is approximately linearly dependent upon the concentration in the concentration range from 0.001 to 0.1M/l (Tamm and Kurtze). The absorption cross section Q is defined by

$$Q = 2\alpha/nL, \tag{87}$$

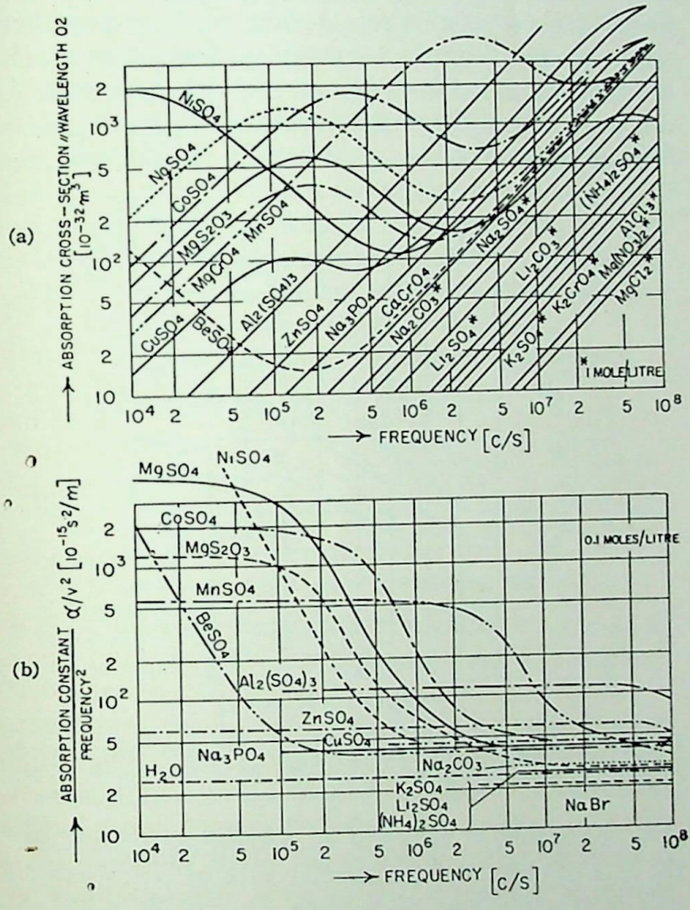


FIG. 8. (a) Sound absorption spectra of electrolytes in aqueous solutions at 20°C according to measurements of Tamm, Kurtze, and Kaiser (relative to water) for 10^{-2} to 10^{-1} M solutions. (b) α/ν^2 for 0.1-M/l solutions of some electrolytes plotted against frequency. [In Fig. 8(a) read $MgSO_4$ for $NgSO_4$ and replace "//" by "×" before and "02" by "$Q\lambda$" after wavelength.]

where n is the concentration in moles per liter and L is the number of molecules per mole. This can be written as

$$Q[m^2] = 1.66\times10^{-25}[2\alpha(\mathrm{cm}^{-1})/n(\mathrm{M/l})]. \tag{88}$$

The absorption cross section is thus found to be constant in the range 0.001 to 0.1 M/l. Figure 9 shows the dependence of the absorption cross section Q of $MgSO_4$, $MnSO_4$, $ZnSO_4$, and $CuSO_4$ on concentration (Tamm and Kurtze[16c]). Wilson carried out the measurements on the concentration behavior of the excess sound absorption in $MgSO_4$ in the concentration range from 0.003 to

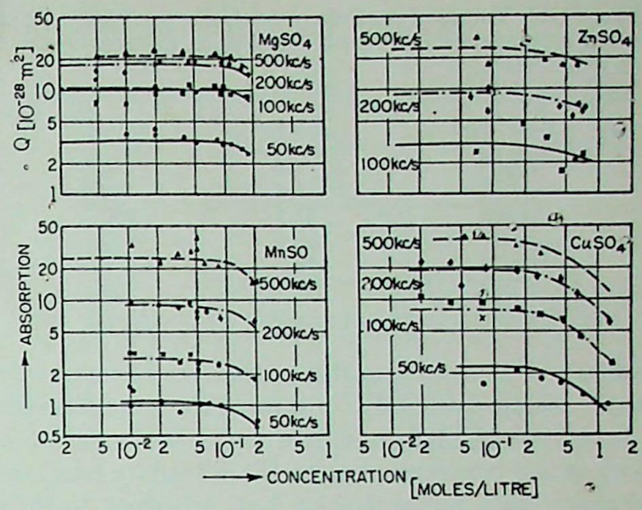


FIG. 9. Dependence of the absorption cross section (Q) of $MgSO_4$, $MnSO_4$, and $CuSO_4$ on concentration.

0.02M. According to him the relationship between the excess sound absorption and the concentration is nonlinear; in fact, the absorption cross section tends to zero as the concentration goes to zero. Table III gives the values of the maximum absorption per wavelength in $MgSO_4$ solutions at several concentrations, temperature being 23° to 25°C.

Kor and Verma[23] have studied the concentration behavior of $MnSO_4$ in great detail from 0.0025 to 1.0M. The relationship between the excess sound absorption and the concentration is found to be linear from 0.02 to 0.1M (Fig. 10), but becomes nonlinear both below 0.02M (Fig. 10) and above 0.1M (Fig. 11). The absorption cross section is constant over the range 0.02 to 0.1M and then diminishes as the concentration increases beyond 0.1M. As the concentration is lowered below 0.02M, the absorption cross section increases.

(c) Effect of Concentration on Relaxation Frequency

According to Wilson's measurements in $MgSO_4$, the relaxation frequency remains approximately constant in the range of concentration from 0.003 to 0.02M.

TABLE II. Values of primary and secondary relaxation frequencies in some of the electrolytes.

Electrolyte	Primary relaxation frequency	Secondary relaxation frequency
$FeSO_4$	10^3 cps	300 Mc
$NiSO_4$	10^4 cps	...
$MgSO_4$	130 kc	200 Mc
MgS_2O_3	180 kc	No relaxation in 100–300 Mc
$MgCrO_4$	180 kc	200 Mc
$CoSO_4$	400 kc	200 Mc
$MnSO_4$	3 Mc	30 Mc
$ZnSO_4$	200 Mc	...
$CuSO_4$	150 kc	>100 Mc
$Al_2(SO_4)_3$	20 Mc (very low conc.)	...
$La(NO_3)_2$	50 Mc	...

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

TABLE III. Values of maximum absorption per wavelength and relaxation frequency in $MgSO_4$ solutions at different concentrations at 23°–25°C.

Conc. M/l	$\mu_m \times 10^6$	ν_m (kc)
0.003	4.6	130
0.0053	15.5	140
0.008	25.5	140
0.01	35.5	130
0.014	48	138
0.02	80	150

Tamm and Kurtze also report that the relaxation frequency is approximately constant in the range from 0.01 to 0.1M. In $MnSO_4$ Smithson and Litovitz report no effect of the concentration on the secondary relaxation frequency with concentration but, according to them, there is indication of a small increase beyond 0.1M. According to Kor and Verma the relaxation frequency is independent of the concentration in the range 0.02 to 0.1M but increases with concentration in the range 0.0025 to 0.02M. It again increases with concentration above 0.1M.

Smithson and Litovitz argue that the specific reaction rate k_1 is concentration dependent too (which has been neglected in the past) and the effect of increasing the ionic strength by increasing the concentration is always to increase the reaction rate, whatever may be the nature of the reaction whether it is between two ions of same sign or opposite signs or between ions and neutral molecules. Hence very small concentration dependence of the reaction rate may favor the reaction between two dipoles. Dissociation of an ion pair also remains a possibility, but at present there is no theory for the effect of ionic strength on the dissociation of an ion pair.

(d) Effect of Dielectric Constant

According to Tamm and Kurtze the electrolytic dissociation is caused by the dipole forces of the molecules of the solute, a measure of this dissociation is the dielectric constant of the liquid. In liquids with smaller dielectric constant than water, the dissociation energy therefore is larger and the relaxation frequencies shifted to lower values. Tamm and Kurtze used water-alcohol mixtures to vary the dielectric constant of the solvent and report lowering of the relaxation frequency with the decrease in the dielectric constant of the solvent. However, the maximum absorption increases in magnitude, the reason for which, according to them, most probably lies in the increase in the percentage of the undissociated molecules.

Bies used dioxane-water mixtures as a solvent. According to his measurements there is small change in the relaxation frequency with concentration. By extrapolating it to zero concentration, he determined the forward reaction rate k_1 on the assumption that the mechanism responsible for the excess sound absorption is of the dissociation type $MgSO_4 \rightleftarrows Mg^{++} + SO_4^{--}$. He found that the reaction rate k_1 first went up and then down as the dielectric constant of the solvent was lowered. He calculated the reverse reaction rate k_2 utilizing the experimentally measured values of equilibrium constant K by conductivity bridge and the

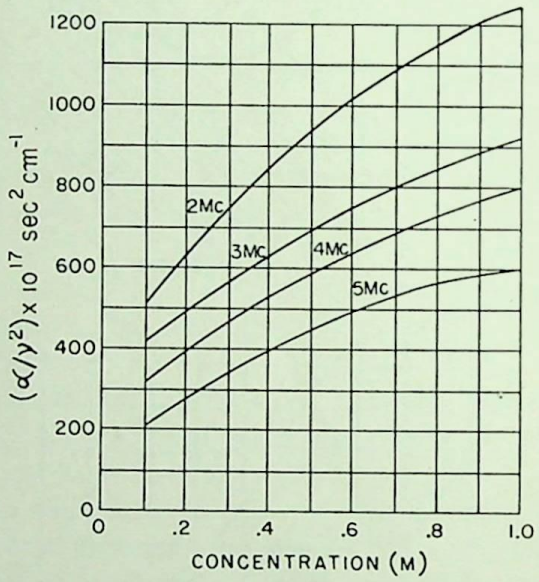


FIG. 11. Nonlinearity of the absorption coefficient with concentration at various frequencies above 0.1 M in $MnSO_4$ at 25°C.

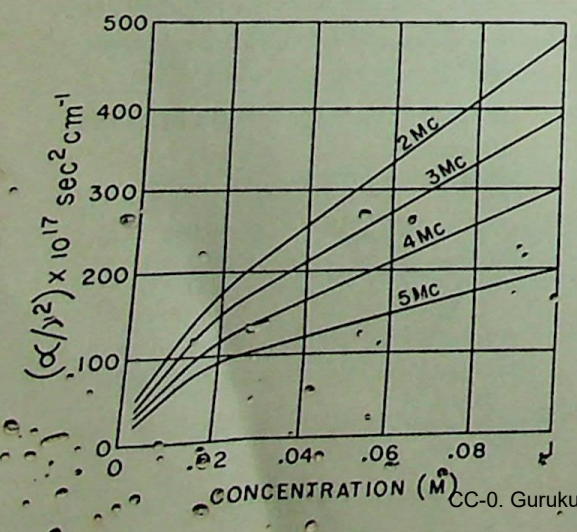


FIG. 10. Variation of the absorption coefficient (α/ν^2) with concentration (0.0025 to 0.1 M) at various frequencies in $MnSO_4$ at 25°C.

above extrapolated values of k_1. According to Laidler,[62] the rate of change of the specific rate of reaction of a pair of reacting ions of valence Z_1 and Z_2 with the reciprocal of the dielectric constant is given by the equation

$$[-d \ln k_2/d(1/D)] = (\epsilon^2 Z_1 Z_2/KTr), \qquad (89)$$

where $Z_1 = -Z_2 = 2$ in the above dissociation, r is equal to twice the mean ionic radius 3.08×10^{-8} cm, K the Boltzmann constant, and ϵ is the electrostatic unit of charge. Bies reported that the specific rate of association does increase with the reciprocal of the dielectric constant of the medium as it should if ions of opposite signs are involved, but apparently not in a linear fashion as expected from the above expression. However, Bies gave no explanation for his observation that the forward reaction rate went first up and then down

[62] K. Laidler, *Chemical Kinetics* (McGraw-Hill Book Company, Inc., 1950).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

as the dielectric constant was lowered. At present there does not seem any explanation for this anomalous behavior of the reaction rate with respect to the dielectric constant. However, Kor and Verma report continuous increase in the reaction rate k_1 with the lowering of the dielectric constant in $MnSO_4$ solutions. They also used water-dioxane mixtures to vary the dielectric constant of the solvent.

Smithson and Litovitz used methanol-water mixtures for varying the dielectric constant. These mixtures have the advantage that water and methanol are similar and the absorption in such mixtures changes only slightly and almost linearly with percentage methanol (Burton),[63] but in this mixture both the molecules are polar and the interaction between the polar constituents of the mixtures may affect the absorption due to main dipole-dipole reaction if it is so. They find a decrease in the primary relaxation frequency by a small amount with decreasing dielectric constant. However, the secondary relaxation peak was observed to increase with decreasing dielectric constant. The magnitude of the absorption increased considerably as D was lowered.

The theory of the influence of the dielectric constant on the rates of reaction, which has experimental evidence in the case of slow reactions has been discussed in several books; see Amis,[64] Laidler,[62] or Glasstone, Laidler, and Eyring.[65] Table IV, presented by the latter authors, shows the effect of dielectric constant and ionic strength on the rates of a reaction. According to Smithson and Litovitz, the lowering of the primary relaxation frequency with the decrease in the dielectric constant suggests that the forward reaction is either between ions of like sign or between dipole molecules. Either of these reactions would be of the first order if one of the reactants were present in great excess and would behave like a dissociation process.

TABLE IV. Effect of dielectric constant and ionic strength on the rates of a reaction.

Reaction type	Effect of increasing dielectric constant	Effect of increasing ionic strength
1. Two dipole molecules	increase	no effect
2. Two ions		
(a) same sign	increase	increase
(b) opposite sign	decrease	increase
3. Ion and neutral molecule	decrease	increase

However, the reaction involving two charged ions of the same sign does not seem to satisfy the condition of large excess. Hence, they conclude in favor of a forward reaction between solute molecules (ion pairs) and the dipole solvent molecules, and the excess of the solvent assures that the reaction is of the first order.

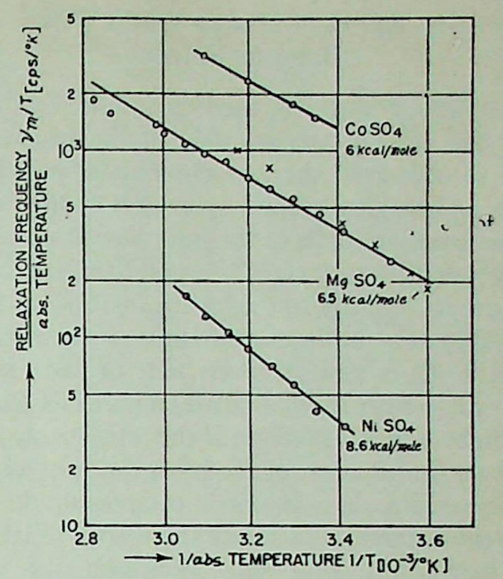


FIG. 12. Temperature dependence of the relaxation frequency: ν_m/T *vs* $1/T$ for $CoSO_4$, $MgSO_4$, and $NiSO_4$.

(e) Effect of Temperature

The relaxation frequency ν_m increases with increase in temperature, while the maximum excess absorption is not altered appreciably. A plot of $\log\nu_m/T$ *vs* $1/T$ gives a straight line, as found by Wilson for $MgSO_4$, by Tamm and Kurtze for $MgSO_4$, $CoSO_4$, and $NiSO_4$ (Fig.12), and by Smithson and Litovitz, and Kor and Verma for $MnSO_4$. The experimental activation energy for the relaxation process obtained from the slope of the above curve are about 6.5 kcal/M for $MgSO_4$, 8.6 kcal/M for $NiSO_4$, about 6 kcal/M for $CoSO_4$. In $MnSO_4$ Smithson and Litovitz found activation energy to be 8.68 kcal/M in the isodielectric solution. Isocomposition measurements were made with two different solvents, pure water and 29.52% methanol gave energies of 7.85 and 8.05 kcal/M, respectively, both of which are lower than isodielectric activation energy.

According to Amis and Jaffe[66] the activation energy ΔE for an ionic reaction consists of two parts, ΔE_1, which represents a contribution to activation energy due to electrostatic forces and ΔE_2, a contribution due to nonelectrostatic forces, $\Delta E=\Delta E_1+\Delta E_2$. According to Amis ΔE_1, in the case of a reaction between two bivalent ions, should be equal to 50% of the total activation energy. However, Smithson and Litovitz, using the methods outlined by Amis and utilizing their values of $d(k_1)/d(1/T)$ and $d(k_1)/d(1/D)$, find ΔE_1 to be just equal to 8% of the total activation energy. Such a low value of the activation energy for the electrostatic contribution, according to them, rules out ion-ion reaction as the rate determining step. However, such a low value may fit with the dipole-dipole reaction.

[63] C. J. Burton, J. Acoust. Soc. Am. **20**, 186 (1948).
[64] E. S. Amis, *Kinetics of Chemical Change in Solution* (The Macmillan Company, New York, 1949).
[65] Glasstone, Laidler, and Eyring, *Theory of Rate Processes* (McGraw-Hill Book Company, Inc., New York, 1941).
[66] E. S. Amis and G. Jaffe, J. Chem. Phys. **10**, 646 (1942).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

(f) Effect of Heavy Water When Used as Solvent

When heavy water is used as a solvent in place of water, the rate of the reaction in the solution is expected to change if the rate-determining step involves making or breaking a bond between a hydrogen and an oxygen atom, apart from the pure solvent effect, which is very small, since the dielectric constant of heavy water differs only by less than 1% and the viscosity by about 20%. The half-quantum zero-point energy in the case of D_2O is smaller than that of H_2O since the heavier deuterium atom oscillates more slowly compared to the lighter protium atom. If the total energy required to make a molecule react is almost the same whether it is a deuterium or a protium compound, the critical increment is larger and hence the reaction rate smaller for a deuterium compound. Smithson and Litovitz, however, notice no measurable shift in the lower relaxation peak. This means that the rate determining step does not involve the OH bond. It rules out such possibilities as Eqs. (69) or (70). However, a secondary relaxation peak is not observed in D_2O solvent. This might be due to the OH bond which might have shifted out of the frequency range or that the relaxation which is normally observed at about 200 Mc in H_2O solution is shifted down so as to obscure the small peak.

(g) Effect of Pressure

Fisher[67] for the first time reported that acoustic absorption in $MgSO_4$ solutions shows a considerable decrease with elevated pressure. Absorption measurements were made near 100 kc by means of a resonant cylindrical steel cavity which could be pressurized. Observations at several frequencies in this region give the pressure dependence of the relaxation frequency. Carnevale and Litovitz[68] have also studied the effect of pressure on ultrasonic relaxation in manganous sulfate and ammonium hydroxide. Absorption data were obtained for $MnSO_4$ at frequencies from 9 to 75 Mc/sec up to pressures of 3930 kg/cm^2 and a temperature of 60°C. The relaxation frequency increased from 19.1 Mc/sec at atmospheric pressure to 30.7 Mc/sec at 3930 kg/cm^2. Absorption measurements in NH_4OH were obtained at frequencies from 15 to 75 Mc/sec up to pressures of 2030 kg/cm^2 and a temperature of 45°C. The relaxation frequency increased from 21.2 Mc/sec at atmospheric pressure to 52.0 Mc/sec at 2030 kg/cm^2. The results are explained on the basis of chemical dissociation.

ACKNOWLEDGMENTS

The author wishes to acknowledge with thanks the helpful discussions he has had with Professor Ernest Yeager. He is especially grateful to Professor K. Bannerji for his valuable suggestions. Special thanks are also due Professor Frank Hovorka for the grant which made possible his stay at Western Reserve University and for the facilities which enabled him to complete this article. Finally, the author also wishes to thank Mr. Jacob Schrope for his help in the preparation of the diagrams.

OTHER REFERENCES ON ULTRASONIC ABSORPTION IN LIQUIDS SINCE 1951

Earlier references found in *Reviews of Modern Physics* are

(a) W. T. Richards, "Supersonic phenomena," Revs. Modern Phys. **11**, 36 (1939);
(b) Markham, Beyer, and Lindsay, "Absorption of sound in fluids," Revs. Modern Phys. **23**, 353 (1951);
(c) S. M. Karim and L. Rosenhead, "The second coefficient of viscosity of liquids and gases," Revs. Modern Phys. **24**, 108 (1952);

and in the following textbooks:

(d) Ludwig Bergmann, *Der Ultraschall und Anwendung in Wissenschaft und Technik* (S. Hirzel Verlag, Stuttgart, 1954), sixth edition;
(e) E. G. Richardson, *Ultrasonic Physics* (Elsevier Publishing Company, London, 1952);
(f) P. Vigoureux, *Ultrasonics* (Chapman and Hall Ltd., London, 1950);
(g) E. G. Richardson, *Technical Aspects of Sound* (Elsevier Publishing Company, London, 1957), Vols. I and II;
(h) A. E. Crawford, *Ultrasonic Engineering* (Academic Press, Inc., New York, 1955);
(i) T. F. Heuter and R. H. Bolt, *Sonics* (John Wiley and Sons, Inc., New York, 1955);
(j) E. G. Richardson, *Relaxation Spectrometry* (Interscience Publishers, Inc., New York, 1957);
(k) K. F. Herzfeld and T. A. Litovitz, *Absorption and Dispersion of Ultrasonic Waves* (Academic Press, Inc., New York, to be published).

Further references include:

(1) T. A. Litovitz, "Ultrasonic absorption of glycerin in the liquid and vitreous state," J. Acoust. Soc. Am. **23**, 75 (1951).
(2) K. Tamm and G. Kurtze, "Absorption of sound in aqueous solutions of electrolytes," Nature **168**, 346 (1951).
(3) E. Freedman, "On the propagation of ultrasonic waves in acetic acid," J. Chem. Phys. **19**, 1318 (1951).
(4) A. G. Chynoweth and W. T. Schneider, "Ultrasonic propagation in binary liquid system near their critical solution temperature," J. Chem. Phys. **19**, 1566 (1951).
(5) R. E. Barrett and R. T. Beyer, "Anomalous effect in the ultrasonic absorption of an electrolytic solution," Phys. Rev. **84**, 1061 (1951).
(6) A. B. Pippard, "Ultrasonic propagation in liquid helium near lambda point," Phil. Mag. **42**, 1209 (1951).
(7) A. N. Hunter, "Ultrasonic absorption in supercooled liquids," Proc. Phys. Soc. (London) **B64**, 288 (1952).
(8) K. F. Herzfeld, "Origin of ultrasonic absorption in liquids," J. Chem. Phys. **20**, 288 (1952).
(9) D. Sette, "On the elastic relaxation in CS_2," Ricerca sci. **22**, 467 (1952).
(10) D. Sette, "Elastic relaxation and structure of liquids, Ricerca sci. **22**, 461 (1952).
(11) P. A. Hudson and M. Eisner, "Debye-Frenkel theory of ultrasonic absorption," Phys. Rev. **85**, 746 (1952).
(12) H. J. McSkimin, "Measurements of dynamic shear viscosity and stiffness of viscous liquids by means of traveling torsional waves," J. Acoust. Soc. Am. **24**, 355 (1952).
(13) Y. Wada and S. Shumbo, "Experimental investigation on the origin of anomalous absorption of ultrasonic waves in liquids," J. Acoust. Soc. Am. **24**, 199 (1952).
(14) L. H. Hall, "Attenuation of sound resulting from ionic relaxation," J. Acoust. Soc. Am. **24**, 704 (1952).
(15) J. Meixner, "General theory of sound absorption in gases and liquids with allowance of transport phenomena," Acustica **2**, 101 (1952).

[67] F. Fisher, J. Acoust. Soc. Am. **29**, 656 (1957); **30**, 442 (1958).
[68] E. H. Carnevale and T. A. Litovitz, J. Acoust. Soc. Am. **29**, 769 (1957); **30**, 610 (1958).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

(16) I. M. Chalatnikov, "Thermal conductivity and sound absorption in helium II," J. Exptl. Theoret. Phys. 23, 21 (1952).
(17) M. Manes, "On the relaxation time of equilibrium system as related to ultrasonic absorption and reaction kinetics," J. Chem. Phys. 20, 1658 (1952).
(18) S. Parthasarthy, Chari, and Srinivasan, "Ultrasonic absorption in various organic liquids at 15 Mcs," Naturwissenschaften 39, 483 (1952).
(19) Nomoto, Kishimoto, and Ikeda, "Ultrasonic absorption in castor oil," Bull. Kobayashi Inst. Phys. Research 2, 72 (1952).
(20) T. A. Litovitz and D. Sette, "Dielectric and ultrasonic relaxation in glycerol," J. Chem. Phys. 21, 17 (1953).
(21) T. Kishimoto and O. Nomoto, "Absorption of ultrasonic waves in CS_2," Bull. Kobayashi. Inst. Phys. Research 2, 63 (1952).
(22) A. K. Dutta, "Ultrasonic absorption and relaxation mechanism," Indian J. Phys. 35, 279 (1952).
(23) D. Sette, "Ultrasonic absorption in liquid mixtures and structural effects," J. Chem. Phys. 21, 558 (1953).
(24) K. Tamm, "Sound absorption measurement in water and in aqueous solutions of electrolytes in the frequency range 5 kc to 1 Mc," Nachr. Akad. Wiss. Göttingen, Math.-physik. Kl. 80, 110 (1952).
(25) Parthasarthy, Srinivasan, and Chari, "Absorption of ultrasonic waves in liquids at 5 Mc from thermal considerations," Nuovo cimento 10, 264 (1953).
(26) Parthasarthy, Chari, and Mahendroo, "Calorimetric determination of the absorption coefficient in organic liquids," J. phys. radium 14, 366 (1953).
(27) Parthasarthy, Chari, and Mahendroo," Determination of ultrasonic absorption coefficient in liquids by a new technique," Z. Naturforsch. 8a, 272 (1953).
(28) C. E. Chase, "Ultrasonic measurements in liquid helium," Phys. Rev. 91, 489 (1953).
(29) F. E. Fox and T. M. Marion, "Ultrasonic dispersion in water solutions of $MgSO_4$," J. Acoust. Soc. Am. 25, 661 (1953).
(30) P. L. Connoly and F. E. Fox, "Ultrasonic absorption in $MgSO_4$ solutions," J. Acoust. Soc. Am. 25, 658 (1953).
(31) M. Pancholy, "Temperature variation of velocity and absorption coefficient of ultrasonic waves in heavy water," J. Acoust. Soc. Am. 25, 1003 (1953).
(32) S. Parthasarthy and A. F. Chapghar," Viscosity as a factor in the anomalous absorption of ultrasonic waves in liquids," Ann. Physik 12, 316 (1953).
(33) I. G. Mikhailov, "Absorption of ultrasonic waves in certain viscous liquids," Doklady Akad. Nauk. S.S.S.R. 89, 991 (1953).
(34) S. Parthasarthy and P. P. Mahendroo, "Absorption coefficient in some liquids determined by a new thermal technique," Nuovo cimento 10, 1196 (1953).
(35) Parthasarthy, Chari, and Srinivasan," Variation of ultrasonic absorption with frequency in organic liquids," Acustica 3, 363 (1953).
(36) I. G. Mikhailov and V. A. Soloviev, "Ultrasonic absorption in liquids and the molecular mechanism of bulk viscosity," Upsekhi Fiz. Nauk. 50, 3 (1952).
(37) M. Manes, "Relationship between kinetics and acoustic phenomenon in equilibrium systems," J. Chem. Phys. 21, 1791 (1953).
(38) E. Freedman, "On the use of ultrasonic absorption for the determination of very rapid reaction rates at equilibrium," J. Chem. Phys. 21, 1784 (1953).
(39) G. F. Alprey and W. G. Schneider, "Ultrasonic absorption in binary liquid systems near the critical region of temperature," Discussions Faraday Soc. 15, 218 (1953).
(40) A. W. Pryor and R. Roscoe, "The velocity and absorption of sound in aqueous sugar solutions," Proc. Phys. Soc. (London) B67, 70 (1954).
(41) Parthasarthy, Srinivasan, and Chari, "The thermal effects of ultrasonic waves in liquids and its relation to their absorption coefficient," Acustica 3, 407 (1954).
(42) C. E. Chase, "Ultrasonic measurements in liquid helium," Proc. Roy. Soc. (London) A220, 116 (1953).
(43) J. F. Mifsud, "Absorption of ultrasonic waves in liquids under pressure," Phys. Rev. 94, 812 (1954).
(44) Litovitz, Higgs, and Meista, "Ultrasonic propagation and its relation to molecular structure in the diols," J. Chem. Phys. 22, 1281 (1954).
(45) Litovitz, Lyon, and Peselnick, "Ultrasonic relaxation and its relation to structure in viscous liquids," J. Acoust. Soc. Am. 26, 566 (1954).
(46) Barret, Beyer, and McNamara, "Ultrasonic absorption in acetate solutions," J. Acoust. Soc. Am. 26, 966 (1954).
(47) E. L. Carstensen, "Relaxational processes in aqueous solutions of $MnSO_4$ and $CoSO_4$," J. Acoust. Soc. Am. 26, 862 (1954).
(48) W. J. Fry and R. B. Fry, "Determination of absolute sound levels and acoustic absorption coefficients by thermocouple probes—experiment and theory," J. Acoust. Soc. Am. 26, 311 (1954); 26, 294 (1954).
(49) F. E. Fox and W. A. Wallace, "Absorption of finite amplitude sound waves," J. Acoust. Soc. Am. 26, 994 (1954).
(50) S. Parthasarthy, "Absorption and dispersion of ultrasonic waves in liquids," J. Acoust. Soc. Am. 26, 611 (1954).
(51) I. G. Mikhailov and L. I. Savina, "The absorption of ultrasonic waves in binary mixtures of liquid with one relaxing component," Doklady Akad. Nauk. S.S.S.R. 96, 1147 (1954).
(52) N. I. Koshkin and V. F. Nozdrev, "Investigation on the absorption of ultrasound in the series of hydrocarbons using the pulse method," Doklady Akad. Nauk. S.S.S.R. 92, 793 (1953).
(53) T. Kishimoto and O. Nomoto, "Absorption of ultrasonic waves in organic liquids I. Liquids with positive temperature coefficient of absorption," J. Phys. Soc. Japan 9, 620 (1954).
(54) N. S. Taskoprulu, "The laws for the excess absorption and the dispersion of sound in $MgSO_4$ solutions at frequencies about 3 Mc," Istanbul Üniv. Fen Fak. Mecmuasi C19, 10 (1954).
(55) E. Grossetti, "Determination of the ultrasonic absorption coefficient of benzol with thermal method," Nuovo cimento 1, 525 (1955).
(56) D. Sette, "Ultrasonic absorption in water-methyl alcohol mixtures," Ricerca. sci. 25, 576 (1955).
(57) D. Sette, "Ultrasonic absorption in aniline-ethyl alcohol mixtures," Acustica 5, 195 (1955).
(58) S. Parthasarthy and A. P. Deshmukh, "Sound absorption in liquids in relation to light scattering data," Ann. Physik 15, 417 (1955).
(59) T. A. Litovitz and E. H. Carnevale, "The effect of pressure on sound propagation in water," J. Appl. Phys. 26, 816 (1955).
(60) C. E. Chase and M. A. Herlin, "Ultrasonic measurements in magnetically cooled liquid helium," Phys. Rev. 99, 669 (1955).
(61) A. Cilesiz, "A new pulse method to measure absorption in transparent liquids," Istanbul Üniv. Fen Fak. Mecmuasi C20, 94 (1955).
(62) R. T. Beyer, "Ultrasonic absorption in toluene," J. Acoust. Soc. Am. 27, 1 (1955).
(63) E. H. Carnevale and T. A. Litovitz, "Pressure dependence of sound propagation in the primary alcohols," J. Acoust. Soc. Am. 27, 547 (1955).
(64) Fox, Culey, and Larson, "Phase velocity and absorption measurements in water containing air bubbles," J. Acoust. Soc. Am. 27, 534 (1955).
(65) M. M. Kannuna, "Measurements of the absorption of ultrasonic waves in liquids by the method of isochromates," J. Acoust. Soc. Am. 27, 5 (1955).
(66) Stakutis, Morse, Dill, and Beyer, "Attenuation of ultrasound in aqueous suspensions," J. Acoust. Soc. Am. 27, 539 (1955).
(67) D. M. Towle and R. B. Lindsay, "Absorption and velocity of ultrasonic waves of finite amplitude in liquids," J. Acoust. Soc. Am. 27, 530 (1955).
(68) E. G. Richardson, "Acoustic experiments relating to the coefficients of viscosity of various liquids," Proc. Roy. Soc. (London) A226, 16 (1954).
(69) Parthasarthy, Chari, and Srinivasan, "The dependence of intensity of sound at source on its absorption coefficient in liquids," J. Chem. Phys. 21, 185 (1953); Physik. Ber. 34, 2031 (1953).
(70) A. W. Pryor and E. G. Richardson, "Velocity and absorption of ultrasonics in liquid sulphur," J. Phys. Chem. 59, 14 (1955).
(71) H. Krishnamuthi, "Relaxational effects in aqueous solutions of sulphur dioxide," Proc. Indian Acad. Sci. 41, 125 (1955).
(72) S. Parthasarthy and M. Pancholy, "Ultrasonic absorption constants in liquids by an improved optical method," Z. Physik 138, 635 (1954); Physik. Ber. 34, 2032 (1954).
(73) Parthasarthy, Tipnis, and Pancholy, "Relation between optically determined absorption of ultrasonic waves and specific heat," Z. Physik 140, 504 (1955).
(74) V. Nozdrev and A. Sultanov, "On the establishment and study of two relaxation regions during the passage of ultrasonic waves in ethyl acetate," Doklady Akad. Nauk. S.S.S.R. 104, 837 (1955).
(75) A. W. Nolle, "The effect of hydrostatic pressure in the two

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

state model of compressional relaxation in liquids," Phys. Rev. **94**, 812 (1954).

(76) S. Parthasarthy and D. S. Guruswamy, "Sound absorption in liquids in relation to their specific heats," Ann. Physik **16**, 31 (1955).

(77) I. Gabrielli and L. Verdini, "Velocity of propagation and absorption coefficient of ultrasonics in mesomorphic liquids," Nuovo cimento **2**, 426 (1955).

(78) T. Lyon and T. A. Litovitz, "Ultrasonic relaxation in normal propyl alcohol," J. Appl. Phys. **27**, 179 (1956).

(79) J. Lamb and D. H. A. Huddart, "The absorption of ultrasonic waves in propionic acid," Trans. Faraday Soc. **46**, 540 (1950); Physik. Ber. **34**, 5836 (1955).

(80) T. Kishimoto and O. Nomoto, "Absorption of ultrasonic waves in organic liquids II. Liquids with negative temperature coefficient of absorption," Bull. Kobayashi Inst. Phys. Research **4**, 175 (1954).

(81) A. Van Itterbeek, "Acoustische Relaxatieversch ynselen in fluida by lage temperaturen," Tijdschr. Natuurk. **21**, 349 (1955).

(82) K. Eppler, "Absorbion of ultrasonics in liquid mixtures," Z. Naturforsch. **10a**, 744 (1955).

(83) B. B. Deo, "Ultrasonic absorption in solutions," Indian J. Phys. **38**, 352 (1955).

(84) C. E. Chase, "Ultrasonic propagation in liquid helium," Am. J. Phys. **24**, 136 (1956).

(85) S. Parthasarthy and A. F. Chapghar, "Sound absorption in liquids in relation to their physical properties—viscosity and specific heats," Ann. Physik **16**, 297 (1955).

(86) S. Parthasarthy and D. S. Guruswamy, "Sound absorption in liquids in relation to their specific heats II," Ann. Physik **16**, 287 (1955).

(87) L. E. Lawley and R. D. C. Reed, "A reverberation method for the measurement of the absorption of ultrasonics in liquids," Acustica **5**, 316 (1955).

(88) T. Kishimoto and O. Nomoto, "Absorption of ultrasonic waves in organic liquids III., Liquids with negative temperature coefficient of sound absorption," J. Phys. Soc. Japan **10**, 933 (1955).

(89) C. R. Rao, "Sound absorption coefficients based on intensity measurements of diffraction orders," Proc. Indian Acad. Sci. **A42**, 331 (1955).

(90) Parthasarthy, Pancholy, and Tipnis, "A new thermal method for sound absorption in liquids," Nature **176**, 611 (1955).

(91) A. Mer and W. Maier, "Absorption of ultrasonics and association in solutions of phenol in CCl_4, C_6H_{12}, and C_6H_5Cl," Z. Naturforsch. **10a**, 997 (1955).

(92) I. G. Mikhailov and G. N. Fep-Fanaov, "Differential method of measuring ultrasonic absorption in fluids," Akust. Z. **2**, 194 (1956).

(93) E. Grossetti, "Determination of the ultrasonic absorption coefficient by thermal methods," Nuovo cimento **3**, 673 (1956).

(94) Bormosov, Nozdrev, Sobolev, and Sultanav, "Experimental investigation of relaxational processes arising in the passage of ultrasonic waves through fluids," Akust. Z. **2**, 118 (1956).

(95) H. J. McSkimin, "Wave propagation and the measurement of the elastic properties of liquids and solids," J. Acoust. Soc. Am. **28**, 1288 (1956).

(96) J. F. Mifsud and A. W. Nolle," Velocity and absorption of ultrasonic waves in several nonassociated liquids under high pressure, J. Acoust. Soc. Am. **28**, 469 (1956).

(97) M. Mokhtar and M. Youssef, "Sensitive electrodynamic balance for measurement of absorption of ultrasonic waves in liquids," J. Acoust. Soc. Am. **28**, 651 (1956).

(98) L. Liebermann, "On the pressure dependence of sound absorption in liquids," J. Acoust. Soc. Am. **28**, 1253 (1956).

(99) V. Narsimhan and R. T. Beyer, "Attenuation of ultrasonic waves of finite amplitude in liquids," J. Acoust. Soc. Am. **28**, 1233 (1956).

(100) M. S. Cohen, "An ultrasonic spectrometer for measuring acoustic absorption," Proc. Natl. Electronics Conf. p. 490 (1956).

(101) S. Parthasarthy and S. S. Mathur, "Ultrasonic absorption in liquids from thermal steady states," Nature **178**, 378 (1956).

(102) S. Parthasarthy and M. Pancholy, "Sound absorption in liquids," Ann. Physik **17**, 417 (1956).

(103) R. Cerf and E. Meyer, "Absorption measurements in liquids by ultrasonic interferometry. Determination of the intrinsic absorption of a macromolecule in solution," Compt. rend. **243**, 148 (1956).

(104) M. S. de Groot and J. Lamb, "Ultrasonic relaxation in unsaturated aldehydes," Nature **177**, 1231 (1956).

(105) Zarembo, Krassilnikov, and Schklovskaya-Kordi, "Absorption of ultrasonic waves of finite amplitude in liquids," Doklady Akad. Nauk. S.S.S.R. **109**, 731 (1956).

(106) J. A. Newell and J. Wilks, "Absorption of sound in liquid helium below 1°K," Phil. Mag. **1**, 588 (1956).

(107) J. H. Andreae and J. Lamb, "Ultrasonic relaxation theory for liquids," Proc. Phys. Soc. (London) **B69**, 814 (1956).

(108) R. T. Beyer, "Ultrasonic absorption—theory of Gierer and Wirtz," J. Chem. Phys. **25**, 219 (1956).

(109)S. Parthasarthy, "Ultrasonic absorption in liquids in relation to their physical properties," Communs. congr. intern. traitments ultrasons, Marseille, p. 71 (May, 1956).

(110) Parthasarthy, Pancholy, and Mathur, "Sound absorption in liquids—thermal methods," Ann. Physik **18**, 220 (1956).

(111) V. F. Nozdrev and V. D. Sobolev, "Investigation of the ultrasonic properties of ethyl acetate in the critical region," Akust. Zhur. **2**, 379 (1956).

(112) O. Nomoto, "Theory of ultrasonic absorption in aqueous solutions I. Introduction and general theory," J. Phys. Soc. Japan **11**, 827 (1956).

(113) J. Lamb, "Ultrasonic relaxation phenomenon in organic liquids," Communs. congr. intern. traitments ultrasons, Marseilles, p. 61 (May, 1956).

(114) A. K. Dutta and K. Samal, "Real and apparent absorption coefficient of ultrasonic waves in liquids," Nature **179**, 95 (1957).

(115) A. N. Hunter, "Measurements of the velocity and absorption of high frequency ultrasonic waves in supercooled liquids," Proc. Phys. Soc. (London) **B69**, 965 (1956).

(116) E. L. Heasell and J. Lamb, "The absorption of ultrasonic waves in a number of pure liquids over the frequency range 100 to 200 Mc/sec," Proc. Phys. Soc. (London) **B69**, 869 (1956).

(117) L. Van Craeynest, "Absorption measurements of ultrasonic waves in water and electrolytic solutions," Verhandel. Koninkl. Vlaam. Acad. Wetenschap. **49** (1955).

(118) G. D. Mikhailov, "On the question of interaction of ultrasonic waves in fluids," Zhur. Eksptl. i Teoret. Fiz. **30**, 1142 (1956).

(119) Sarembo, Krasilinikov, and Skhlovskaya-Kordi, "On the propagation of ultrasonic large amplitude waves in liquids," Akust. Z. **3**, 29 (1957).

(120) R. T. Beyer, "Recent research in ultrasonics & physical acoustics in U.S.S.R.," Nuovo cimento **4**, 31 (1956).

(121) F. A. Levi, "On the measurement of ultrasonic intensities by means of a radiation pressure balance," Nuovo cimento **4**, 1073 (1956).

(122) S. Parthasarthy and P. P. Mahendroo, "Relation between efficiency of quartz transducers and ultrasonic absorption coefficients of liquids. I, II.," Z. Physik **147**, 573 (1957).

(123) S. Parthasarthy and S. S. Mathur, "Ultrasonic absorption by steady thermal methods," Ann. Physik **19**, 242 (1956).

(124) O. Nomoto, "Theory of the ultrasonic absorption in aqueous solutions II. Aqueous solutions of some alcohols," J. Phys. Soc. Japan **12**, 300 (1957).

(125) W. Maier and H. D. Rudolph, "Ultrasonic absorption measurements in dilute solutions for determining the rates of association and dissociation of benzoic acid in CCl_4," Z. physik. Chem. **10**, 83 (1957).

(126) T. A. Litovitz, "Theory of ultrasonic thermal relaxation in liquids," J. Chem. Phys. **26**, 469 (1957).

(127) Litovitz, Carnevale, and Kendall, "Effects of pressure on ultrasonic relaxation in liquids I," J. Chem. Phys. **26**, 465 (1957).

(128) J. H. Andreae, "Ultrasonic relaxation in methylene chloride," Proc. Phys. Soc. (London) **B70**, 71 (1957).

(129) R. Piccirelli and T. A. Litovitz, "Ultrasonic shear and compressional relaxation in liquid glycerol," J. Acoust. Soc. Am. **29**, 1009 (1957).

(130) H. J. McSkimin, "Ultrasonic pulse technique for measuring acoustic losses and velocities of propagation in liquids as a function of temperature and hydrostatic pressure," J. Acoust. Soc. Am. **29**, 1185 (1957).

(131) T. Tashimoto and Y. Kikuchi, "Ultrasonic propagation measurement in sea water up to 400 Kc," J. Acoust. Soc. Am. **29**, 702 (1957).

(132) T. A. Litovitz, "Relation of nuclear spin-lattice relaxa-

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

tion to ultrasonic viscous relaxation in liquids," J. Acoust. Soc. Am. 29, 648 (1957).
(133) Krassilnikov, Shklovskayakordy, and Zarembo, "On the propagation of ultrasonic waves of finite amplitude in liquids," J. Acoust. Soc. Am. 29, 642 (1957).
(134) K. F. Herzfeld, "On the origin of ultrasonic absorption in liquids," J. Acoust. Soc. Am. 29, 1180 (1957).
(135) Schuele, Gutowski, and Carome, "Interferometric determination of ultrasonic absorption in castor oil," J. Acoust. Soc. Am. 29, 1081 (1957).
(136) R. T. Beyer, "Double relaxation effects," J. Acoust. Soc. Am. 29, 247 (1957).
(137) D. Tabuchi, "Dispersion and absorption of sound in liquids in general chemical equilibrium and its application to chemical kinetics," J. Chem. Phys. 26, 993 (1957).
(138) L. K. Sarembo, "On the temperature dependence of high intensity sound absorption in viscous liquids," Akust. Zhur. 3, 163 (1957).
(139) V. F. Nozdrev and V. D. Sobolev, "Measurement of ultrasonic wave absorption in ethyl acetate at the saturation line by a pulse method with a double fixed distance," Doklady Akad. Nauk. S.S.S.R. III, 808 (1956).
(140) R. E. Nettleton, "Theory of the equation of state and ultrasonic absorption in associated liquids at ordinary temperatures," Phys. Rev. 106, 631 (1957).
(141) I. G. Mikhailov, "Ultrasonic absorption in viscous liquids," Akust. Zhur. 3, 177 (1957).
(142) O. Nomoto, "Phenomenological theory of the molecular absorption and dispersion of sound in fluids and the relation between the relaxation time of the internal energy and the relaxation time of the internal specific heat," J. Phys. Soc. Japan 12, 85 (1957).
(143) A. Carreli and F. S. Gaeta, "A new method for the determination of the acoustic absorption coefficient in liquids," Nuovo cimento 5, 773 (1957).
(144) R. I. Tait, "An apparatus for the study of ultrasonic propagation in liquids under pressure," Acustica 7, 193 (1957).
(145) E. G. Richardson and R. I. Tait, "Ratio of specific heat and high frequency viscosities in organic liquids under pressure derived from ultrasonic propagation," Phil. Mag. 2, 441 (1957).
(146) "Molecular aspects of sound propagation in gases and liquids," Proc. Intern. Comm. Acoust. Congr. 2nd. Cambridge Massachusetts, 1956 (published 1957).
(147) R. O. Davies and J. Lamb, "Ultrasonic analysis of molecular relaxational processes," Chem. Soc. London Quart. Revs. 11, 134 (1957).
(148) W. Craya, "Two fluid model and sound absorption in helium II," Helv. Phys. Acta 30, 447 (1957).
(149) V. P. Glotov, "The theory of the relaxation absorption and dispersion of sound in strong electrolytes which are not fully dissociated," Soviet Phys. Acoustics 3, 236 (1957).
(150) G. Laville, "Apparatus for measuring the absorption of ultrasound in liquids," Compt. rend. 245, 1523 (1957).
(151) M. S. de Groot and J. Lamb, "Ultrasonic relaxation in the study of rotational isomers," Proc. Roy. Soc. (London) A242, 36 (1957).
(152) O. Nomoto, "Ultrasonic absorption and dispersion in liquids in relation to liquid structure II. Liquid mixtures and solutions," J. Acoust. Soc. Japan 13, 156 (1957).
(153) R. S. Musa, "Two crystal interferometric method for measuring ultrasonic absorption coefficient in liquids," J. Acoust. Soc. Am. 30, 215 (1958).
(154) T. A. Litovitz, "Origin of ultrasonic volume viscosity in associated liquids," J. Acoust. Soc. Am. 30, 210 (1958).
(155) Dransfeld, Newell, and Wilks, "The absorption of sound in liquid helium II," Proc. Roy. Soc. (London) A243, 500 (1958).
(156) M. Mokhtar and H. Youssef, "Ultrasonic absorption in liquids I, J. Acoust. Soc. Am. 30, 549 (1958).
(157) D. Tabuchi, "Dispersion and absorption of sound in ethyl formate and study of the rotational isomers," J. Chem. Phys. 28, 1014 (1958).
(158) R. M. Mazo, "Absorption and dispersion of sound in chemically reacting fluid," J. Chem. Phys. 28, 1223 (1958).
(159) N. Hirai and H. Eyring, "Bulk viscosity of liquids," J. Appl. Phys. 29, 810 (1958).
(160) M. Cevolani and S. Petralia, "Ultrasonic absorption in aniline and in mixtures," Nuovo cimento 7, 866 (1958).
(161) K. Herzfeld, "Bulk viscosity and shear viscosity in fluids according to the theory of irreversible processes," J. Chem. Phys. 28, 595 (1958).
(162) S. Parthasarthy and M. Pancholy, "Ultrasonic velocity and absorption coefficient in a chemically reactive medium," Z. angew. Phys. 10, 193 (1958).
(163) H. Singh, "Sound absorption in relation to free volume of liquids," Nuovo cimento 9, 545 (1958).
(164) B. R. Ramachandra Rao and H. S. Rama Rao, "A method of studying the variation with temperature of ultrasonic absorption in liquids," Nature 182, 1794 (1958).
(165) S. Parthasarthy and M. Pancholy, "Studies in ultrasonic propagation in mixtures of ethyl alcohol and water," Z. angew. Phys. 10, 453 (1958).
(166) P. White and G. C. Benson, "Ultrasonic absorption in (pure) butyric acid," Can. J. Chem. 36, 1135 (1958).
(167) Murphy, Ganison, and Potter, "Sound absorption at 50 to 500 kc from transmission measurements in the sea," J. Acoust. Soc. Am. 30, 871 (1958).
(168) K. Barron and L. E. Lawley, "Absorption of sound in sediments of small particles in liquids," Proc. Phys. Soc. (London) 72, 933 (1958).
(169) R. Bass and J. Lamb, "Ultrasonic relaxation of the vibrational specific heat of carbon dioxide, sulphur hexafluoride, nitrous oxide, cyclopropane and methyl chloride in the liquid state," Proc. Roy. Soc. (London) A247, 168 (1958).
(170) R. T. Beyer and V. Narsimhan, "On the absorption of ultrasonic waves of finite amplitude in liquids," Soviet Phys.-Acoust. 4, 196 (1958).
(171) V. P. Glotov, "Reverberation tank method for the study of sound absorption in the sea," Soviet Phys.-Acoust. 4, 243 (1958).
(172) B. I. Kal'ianov and V. F. Nozdrev, "Investigation of the frequency-temperature dependence of the ultrasonic absorption coefficient in the critical region of methyl acetate," Soviet Phys.-Acoust. 4, 198 (1958).
(173) I. G. Mikhailov, "On the absorption of ultrasonic waves in ethyl acetate," Soviet Phys.-Acoust. 4, 200 (1958).
(174) V. F. Nozdrev, "On the absorption of ultrasonic waves in ethyl acetate," Soviet Phys.-Acoust. 4, 205 (1958).
(175) Parthasarthy, Pancholy, and Chapghar, "Ultrasonic absorption in some homologous series of organic liquids," Nuovo cimento 10, 111 (1958).
(176) Y. Wada, "Attenuation of ultrasonic waves in suspensions of viscoelastic spheres," J. Phys. Soc. Japan 13, 1390 (1958).
(177) R. E. Nettleton, "Compressional relaxation in liquids," J. Acoust. Soc. Am. 31, 557 (1959).
(178) R. O. Davies, "Some equilibrium and non-equilibrium properties of an $(s+1)$-state ideal solution," Proc. Phys. Soc. (London) 73, 25 (1959).
(179) R. O. Davies and J. Lamb, "On the derivation of reaction parameters from observations of ultrasonic relaxation," Proc. Phys. Soc. (London) 73, 767 (1959).
(180) D. N. Hall and J. Lamb, "The effect of polar and non-polar solvents on the isomeric reaction in ethyl formate demonstrated by ultrasonic measurements," Trans. Faraday Soc. 55, 794 (1959).
(181) T. A. Litovitz, "Ultrasonic spectroscopy in liquids," J. Acoust. Soc. Am. 31, 681 (1959).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

Anisotropic Light Scattering of Streaming Suspensions and Solutions

WILFRIED HELLER

Chemistry Department, Wayne State University, Detroit 1, Michigan

I. INTRODUCTION

IN recent years considerable use has been made of light scattering for determining the shape of dissolved macromolecules and of dispersed colloidal particles. The favored method uses the variation of scattering with the angle of observation. The theories used are due to Gans[1] and Debye.[2] An alternative method is that of Krishnan.[3] Here, the relative intensities of the "four Krishnan components" of light scattered at 90° with respect to the primary beam are the experimental data. The former method uses incident natural light; the latter, incident plane polarized light. Both have in common the investigation of light scattering by systems with randomly oriented molecules or particles. There is no need for a review of either of these methods since several excellent summaries are already available.[4] A much neglected, yet more powerful and older third method, is due to Diesselhorst and Freundlich.[5] In this, the light scattering is observed subsequent to an orientation of the scattering bodies, e.g., by streaming. Such an orientation produces an anisotropy of light scattering: the scattered intensity varies with the orientation of the electric vector of the incident beam with respect to the morphologically extraordinary direction of the scattering bodies. This review is concerned with the three principal phenomena of anisotropic light scattering observed or to be expected: dityndallism, conservative dichroism, and "bidissymmetry."

II. METHODS AND PROBLEMS OF ORIENTATION

The necessary orientation of particles or molecules may be produced by streaming, or by an electric or magnetic field. Flow orientation is due exclusively to anisometric (nonspherical) shape and is therefore of primary interest here. It is "kinematic"—although the particles rotate continuously due to the velocity gradient q and to the rotatory Brownian movement, they spend most of their time in a position defined by the angle χ. This is the angle between the most probable direction of the longest dimension of the particles and the direction of the velocity gradient. If the particles or molecules are deformable, the phenomenon is complicated by dilation at an angle of 45° with respect to q and the direction of flow, and compression at an orthogonal angle. The direction of flow is generally referred to as the extraordinary direction e of the system. It will actually be an extraordinary direction only if $q \rightarrow \infty$, i.e., when the longest dimension of the particles is parallel to the direction of flow ($\chi = 90°$). If the particles are uniaxial, the direction of flow becomes then—according to the definition in crystal optics—the optic axis of the system and, therefore, the optically extraordinary direction.

Magnetic and electric orientation are simpler in nature since the direction of orientation always coincides with the direction of the field or with the direction normal to it. The degree of orientation is governed here by the torque acting upon the particles and the superimposed randomizing effect of the rotatory Brownian movement. Complications arise from the fact that magnetic and electric orientation may be due to both anisometry and intrinsic (magnetic or electric) anisotropy. Since the orienting torque due to anisometry is generally very weak the orientation is, in general, mainly due to a permanent or induced magnetic or electric dipole. The existence of the two torques can lead to considerable complication if they are orthogonal. The orientation may then change direction with field strength, temperature, or frequency. Some of the resulting optical complications are described in detail elsewhere.[6] Of particular interest is the application of alternating magnetic[7] or electric[8] fields. Whenever magnetic orientation is applicable, it deserves preference over electric orientation because of a variety of possible secondary complications of the latter.[8]

III. DITYNDALLISM

1. Classical Experiments and Definitions

Dityndallism was discovered by Diesselhorst and Freundlich[5] on various streaming colloidal solutions.

[1] R. Gans in *Handbuch der Experimentalphysik* (Akademische Verlagsgesellschaft, Leipzig, 1928), Vol. 19, p. 34; Ann. Physik 62, 331 (1920).

[2] P. Debye, lecture given at the Polytechnic Institute of Brooklyn, Brooklyn, New York (November 25, 1944).

[3] R. S. Krishnan, Proc. Indian Acad. Sci. A1, 211 (1934) and numerous later papers by this author and his associates.

[4] See, e.g., H. C. van de Hulst, *Light Scattering by Small Particles* (John Wiley & Sons, Inc., New York, 1957).

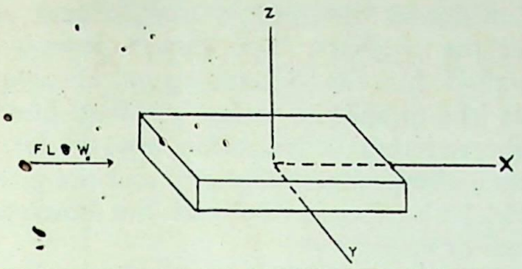
[5] H. Diesselhorst and H. Freundlich, Physik. Z. 16, 422 (1915); 17, 117 (1916).

[6] W. Heller, "Nouvelles recherches sur les propriétés magnéto-optiques des solutions colloidales" in *Actualités sci. et ind.* (Hermann et Cie., editors, Paris, 1939), Vol. 806.

[7] Selected references may be found in W. Heller, *Kolloidphysik des Eisenoxydsols*; in Abegg-Koppel, *Handbuch der anorganischen Chemie* (S. Hirzel, Leipzig, 1934), Vol. 4, Chap. 3, p. 2B.

[8] W. Heller, Revs. Modern Phys. 14, 390 (1942).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA



DIRECTION OF PRIMARY BEAM	DIRECTION OF ELECTRIC VECTOR	DIRECTION OF OBSERVATION	SCATTERED INTENSITY RELATIVE RODS	SCATTERED INTENSITY RELATIVE PLATES
X	Y	Z	●	○
X	Z	Y	●	●
Z	X	Y	○	○
Z	Y	X	●	○
Y	Z	X	●	●
Y	X	Z	○	○

FIG. 1. Dityndallism produced by flow orientation of colloidal rods and plates. ●, scattered intensity is strong. ○, scattered intensity is weak.

Figure 1 illustrates their procedure and results. The colloidal solution flows through a rectangular cell in the x direction. The principal velocity gradient is in the z direction and a secondary gradient is in the y direction. Consequently, rod-like particles will, at a sufficiently high velocity gradient, direct their morphologically extraordinary direction, i.e., the major axis, preferentially toward x, while disk-like particles will direct theirs, the shortest dimension, preferentially toward z. The direction of the linearly polarized beam and of its electric vector and the direction of observation (always perpendicular to the incident beam) are varied to the fullest possible extent.

The black and white disks indicate that the observed scattered intensity is weak or strong, respectively.[9] It follows that, for one particle, the intensity of light scattered at 90° with respect to the incident beam is strongest if the electric vector is perpendicular to the plane of observation and parallel to the largest dimension of the particle. The laterally observed Tyndall beam of a dispersed system of oriented rods has therefore maximal intensity if the electric vector of the incident beam vibrates, for $q \to \infty$, parallel to the direction of flow and has minimal intensity for the orthogonal direction of vibration. This effect, called "dityndallism," [10a] could be defined in analogy to birefringence and dichroism by

$$(D_o)_\theta = (I_e - I_o), \qquad (1)$$

where I is the intensity of light scattered at an angle θ, per unit solid angle, per unit volume of unit concentration, and per unit intensity of the incident beam, and o refers to the ordinary direction of the system. The usefulness of this definition is limited, however, since $(I_e - I_o)$ may be finite also in the unoriented system. An alternate definition,

$$(D)_\theta = (I_e' - I_e'') - (I_o' - I_o'') = \Delta I_e - \Delta I_o, \qquad (1a)$$

therefore may be introduced as preferable. The superscripts $'$ and $''$ refer to the oriented and unoriented system, respectively. For practical work, definition (1a) may be expressed in terms of more convenient intensity ratios.[10b]

Flow dityndallism is an exclusive form effect only if produced by deformation of isotropic fluid material

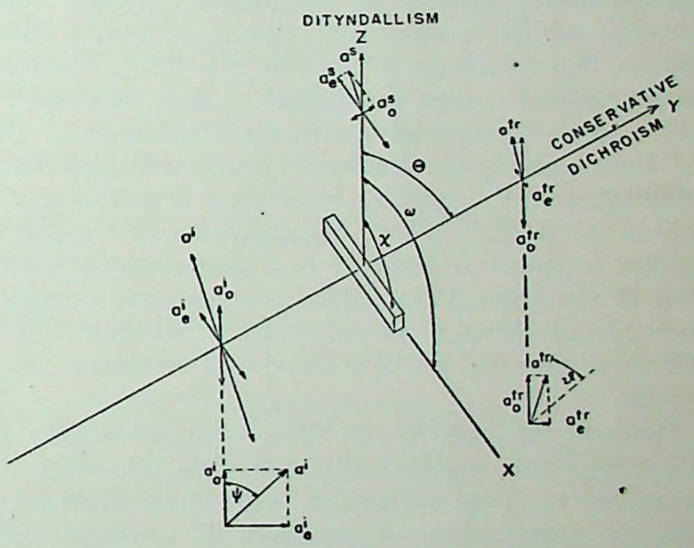


FIG. 2. Geometrical definitions pertinent to flow dityndallism and conservative flow dichroism.

(e.g., emulsion droplets) or by orientation of anisometric intrinsically isotropic bodies. Anisometric bodies are rarely intrinsically isotropic. In general, therefore, an intrinsic dityndallism of equal or opposite sign will be superimposed upon the form dityndallism. A quantitative differentiation between the two effects should be possible by varying the refractive index of the medium.

2. Physical Optics of Dityndallism

The physical optics of dityndallism *is* somewhat intricate. This is discussed with *the aid* of Fig. 2, where the principal drawing is *in perspective*, while two auxiliary drawings pertain *to the xz* plane viewed in the direction of the light *source*. The symbols x, y, and z indicate the direction *of flow* of the primary beam and

[9] A quantitative investigation can be expected to show that there are four different intensities for platelets and three for rodlets.

[10a] H. Zocher, Kolloid Z. 37, 336 (1925).

[10b] M. *Nakagaki* and W. Heller, J. Polymer Sci. (June, 1959).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

of the velocity gradient, respectively. The angle χ between the major axis of the rod-like particle considered and the velocity gradient is, for simplicity, assumed to be 90°. This corresponds to $q \rightarrow \infty$. The angle between the electric vector of the incident beam and the yz plane is ψ, that between the vector of the scattered beam—or of the scattered component considered—and the xz plane is ψ' (symbol not shown in Fig. 2). The angle between the variable plane of observation and the direction of flow is ω, that between the direction of observation and the direction of the primary beam is θ.

Consider first the situation pictured in Fig. 2, in which $\theta=\omega=\chi=90°$ and let the angle ψ be 90° and 0° in two *consecutive* measurements. The amplitude of the component of the incident beam vibrating in the extraordinary direction a_e^i gives rise to an amplitude of the scattered component a_e^s, and, in the more general case, of a component $a_o^{s'}$. The latter is not considered in the drawing. Similarly, a_o^s (and a neglected component $a_e^{s'}$) will result from a_o^i.[11] Since $a_e^s \neq a_o^s$, the consecutively determined amplitudes of the components of the transmitted beam, $a_e^{tr} \neq a_o^{tr}$. Next assume that $\theta=\omega=\chi=90°$, but $(0° \neq \psi \neq 90°)$. The scattering process is then equivalent to that expected for two coherent components with the amplitudes a_e^i and a_o^i. Consequently, the scattered linearly polarized beam, of amplitude a^s, resulting from their interference, vibrates at an angle $\psi' \neq \psi$. Measurement of the dityndallism is therefore reduced to a single measurement, that of the angle ψ' for which the scattered intensity transmitted through an analyzer is maximal.[12] The difference between ψ and ψ' reaches a maximum when $\psi=45°$.

When $\theta=90°$, $\omega \neq 90°$, $\chi=90°$ (and $0° \neq \psi \neq 90°$), the scattered beam is elliptically polarized, the ellipticity being due to phase differences between the orthogonal coherent components of wavelets of scattered light originating in different volume elements of the particle. The situation is the same when $\theta=\omega=90°$, $\psi=0°$, but $\chi \neq 90°$.

This discussion allows one to anticipate a few physical phenomena which have yet to be verified experimentally. It should be possible to determine χ by rotating the direction of observation, in the xz plane, to the particular $\omega \neq 90°$ value at which the ellipticity vanishes, i.e., at which the scattered beam represents linearly polarized light. The rate of change of the state of polarization during this rotation should vary with the shape of the particles. This should provide an auxiliary method for the determination of shapes. Assume that the proper ω value yielding linearly polarized light is found: a transition from $\theta=90°$ to $\theta \neq 90°$ will then also produce ellipticity and the rate of its change with θ also should vary with the dimensions of the particles or molecules.

The case may arise that ellipticity is observed even for $\theta=\omega=90°$, $\chi=90°$, and $0° \neq \psi \neq 90°$. This would be indicative of intrinsic anisotropy and the degree of ellipticity would be indicative of the magnitude of intrinsic anisotropy. Ellipticity would be due here to the difference in retardation of the orthogonal components of scattered light within the scattered particle by virtue of differences in its refractive indices parallel and perpendicular to the rod axis. This effect would represent the true complement, in the scattering process, to intrinsic birefringence.

3. Theory of Dityndallism

The theory of dityndallism is concerned with the scattering process of nonspherical particles and with the distribution function of their principal axes. The latter was treated first by Langevin[13] for the relatively simple case of magnetic (or electric) orientation. This theory yields the relationship

$$(\mu_e-\mu_u)(\mu_u-\mu_o)=2.0, \qquad (2)$$

where the subscript u refers to the unoriented system and μ is the refractive index. The relationship should, of course, hold also for turbidity and true absorption. It was verified for the former effect.[14] The distribution function for electric orientation of flexible macromolecules was derived, more recently, by Isihara.[15] The more complicated case of flow orientation of rigid spheroids has been treated by Boeder[16] and, more rigorously, by Peterlin and Stuart.[17] Finally, the distribution function for the end-to-end distance of flow deformed molecules has been developed recently.[18]

The theory of anisotropic scattering for the case in which intrinsically isotropic and rigid anisometric particles are small compared to the wavelength of the incident beam is implicitly contained in an early treatment by Rayleigh.[19] A recent extension of this theory by A. F. Stevenson includes the dimensions commonly encountered with proteins and synthetic macromolecules. This theory will be published in the near future.

---

[11] Owing to the interference of each of the two disregarded components *with* the respective coherent orthogonal components, a linearly polarized beam results in each instance which vibrates in reality at an *angle* $\psi' \neq 0°$ and $\psi' \neq 90°$, respectively. It is surprising that in depolarization measurements on resting solutions by Krishnan's method *no attention* has been paid either to the existence of these two linearly polarized scattered beams or to their measurement.

[12] Similarly V_v+V_h and H_v+H_h could be determined in one measurement and the four "Krishnan-components" V_v, V_h, H_v, and H_h in two measurements instead of the four generally used in depolarization measurements on unoriented systems.

[13] P. Langevin, Radium **7**, 249 (1910).

[14] W. Heller and G. Quimfe, Phys. Rev. **61**, 382 (1942).

[15] Isihara, Koyama, Yamada, and Nihioka, J. Polymer Sci. **17**, 341 (1955); see also C. Wippler; J. Polymer Sci. **23**, 199 (1957).

[16] P. Boeder, Z. Physik **75**, 258 (1932).

[17] A. Peterlin and H. A. Stuart, *Hand- und Jahrbuch d. Chem. Physik*, A. Eucken and K. L. Wolff, editors (Akademische Verlagagesellschaft, Leipzig, 1943), Vol. VIII, Sec. IB.

[18] Peterlin, Heller, and Nakagaki, J. Chem. Phys. **28**, 470 (1958).

[19] Lord Rayleigh, Phil. Mag. **41**, 447 (1871); **44**, 28 (1897).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

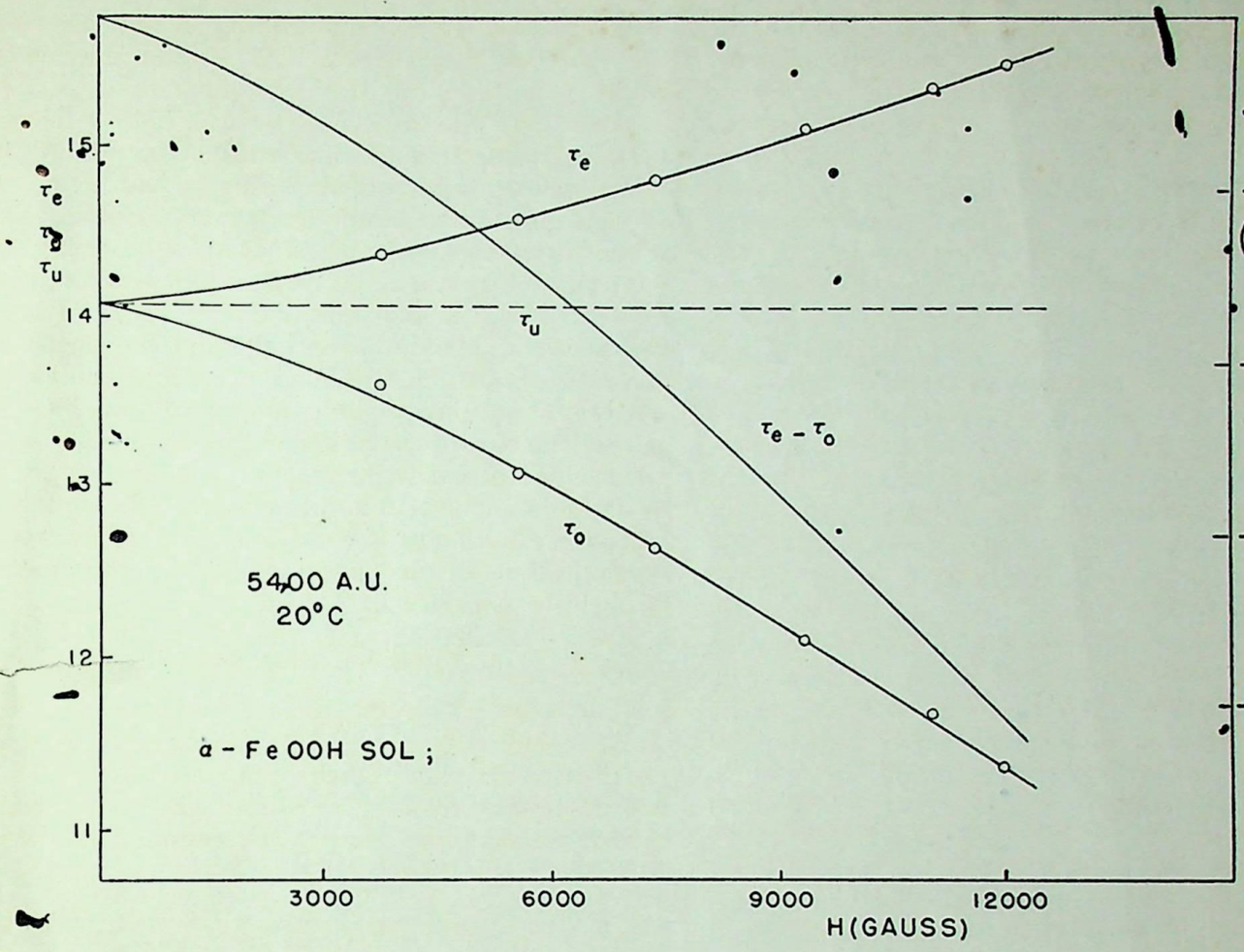


FIG. 3. Conservative magnetic dichroism of an α-FeOOH-sol.

The treatment of anisotropic light scattering to be expected from random coils deformed by flow is difficult, and suitable approximations are therefore necessary. Exploratory investigations to that effect have been published elsewhere.[20,10b]

IV. THE CONSERVATIVE DICHROISM

1. Definitions and Methods of Measurement

Figure 2 shows that the turbidity τ of a dispersed system of oriented anisometric particles varies with the direction of the electric vector of the incident linearly polarized beam. Consequently, the difference $(\tau_e - \tau_o) \neq 0$.[21] The existence of this effect was explicitly proved only rather recently.[6] It was subsequently designated as "conservative" dichroism.[22] This effect, like dityndallism, is primarily an effect of anisometry, but again the intrinsic anisotropy may make a contribution of equal or opposite sign. If the form effect predominates, then the sign of the total effect immediately indicates the orientation of the longest particle dimension with respect to the extraordinary direction. In order to evaluate the relative contribution of the intrinsic effect, one can—as with birefringence—eliminate the form effect by suitably varying the refractive index of the medium.[23]

Investigation of the conservative dichroism requires that one work in a spectral range of negligible absorption. Otherwise, the effect observed is complex in nature because of the simultaneous presence of true dichroism which is due to a difference in the extinction coefficients ϵ_e and ϵ_o. The two effects are physically not separable.

The theory of conservative dichroism, like that of dityndallism, is implicitly contained in the paper by Rayleigh[19] when the particles are small compared to the wavelength. The theory for larger particles is, at present, under preparation. The latter case is of far greater importance since the smallness of the turbidity in systems with very small particles makes it difficult

[20] A. F. Stevenson and H. L. Bhatnagar, J. Chem. Phys. 29, 1336 (1958).

[21] "Turbidity rations" may often be easier to obtain or may convey supplemental information. Thus, the ratios τ_e/τ_u or τ_o/τ_u, where u refers to the unoriented system, give the turbidity of the oriented system, for the particular direction of the electric vector, relative to that of the isotropic unoriented system.

[22] See reference 11 in Heller, Quimfe, and Yeou-Ta, Phys. Rev. 62, 479 (1942).

[23] Results by A. Piekara and W. Heller (to be published).

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

11245

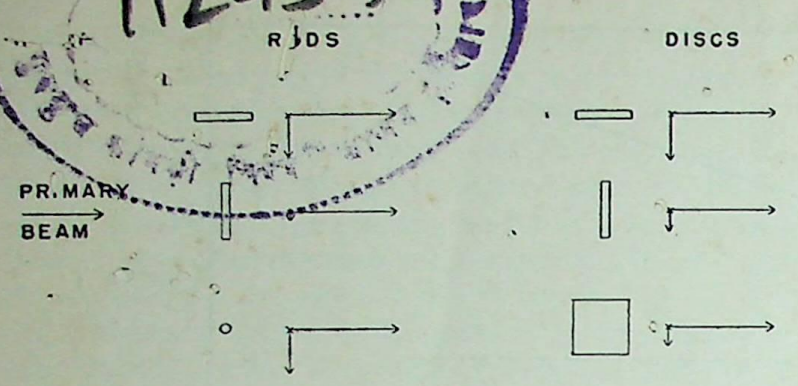


FIG. 4. Anisotropic interference of light scattered by oriented rods and plates not small compared to the wavelength of the incident beam.

to obtain a reliable turbidity difference $(\tau_e - \tau_o)$. Although the attractive feature of conservative dichroism, like that of dityndallism, is primarily the possibility of obtaining numerical data on the dimensions of nonspherical bodies, other applications also are interesting: one can easily obtain semiquantitative information on changes in size, shape, and structure of aggregates formed during coagulation of anisometric particles.[24] Furthermore, the mechanism and kinetics of sol-gel transformation and the change in rotatory mobility of particles during this process can be studied successfully.[25]

Several methods are available for experimental determination of $(\tau_e - \tau_0)$. First, by conventional spectrophotometric procedure τ_e and τ_0 may be measured in succession. The plane of polarization is rotated, between measurements, by 90° and the oriented system and the reference medium (solvent) are brought into the light path consecutively. In the case of magnetic orientation, one may use a split beam of orthogonally polarized components, each of which traverses one compartment of a twin cell filled with the oriented system and solvent, respectively. The second measurement is made after rotating the twin cell by 180° about an axis perpendicular to the field and incident twin beam.[26] Figure 3 shows the magnetic conservative dichroism thus obtained in an old colloidal solution of FeOOH.[27] (The results, to which the true dichroism makes a negligible contribution at 5400 a.u. show immediately that the plate-like particles of submicroscopic size orient themselves parallel to the lines of force.) These two methods give τ_u, in addition to τ_e and τ_o. This is not the case for the third method, which is the *fastest*, and also the most sensitive if the effects are small. *Here*, $\psi = 45°$ (Fig. 2). The conservative dichroism then produces a rotation of the plane of polarization ϑ[28] from which $(\tau_e - \tau_o)$ is easily derived.[29]

V. BIDISSYMMETRY IN UNPOLARIZED LIGHT

The rods and disks, shown schematically in Fig. 4 in the three possible orientations with respect to an incident unpolarized beam, lead to radiation diagrams of which two vectors are shown schematically in each case: the radius vector at 90° and that at 0° with respect to the direction of the primary beam. The former vector applies to observation in the plane of the paper. The reduction, by interference, in the intensity of laterally scattered light is more pronounced if the longest direction of the particle is perpendicular to the primary beam and is directed towards the observer. Again, a differentiation between rods and disks should be possible, although the quantitative differences in the optical effects can be expected to be smaller than for incident polarized light.[30] Intrinsic anisotropy also makes a contribution here, since interference varies with the refractive index.

Of particular interest is the fact that here light scattering exhibits a twofold dissymmetry. This is shown schematically in Fig. 5. The cross sections of the schematic radiation diagrams of spheres, random coils, or randomly oriented nonspherical bodies (left) and of oriented nonspherical bodies or deformed random coils

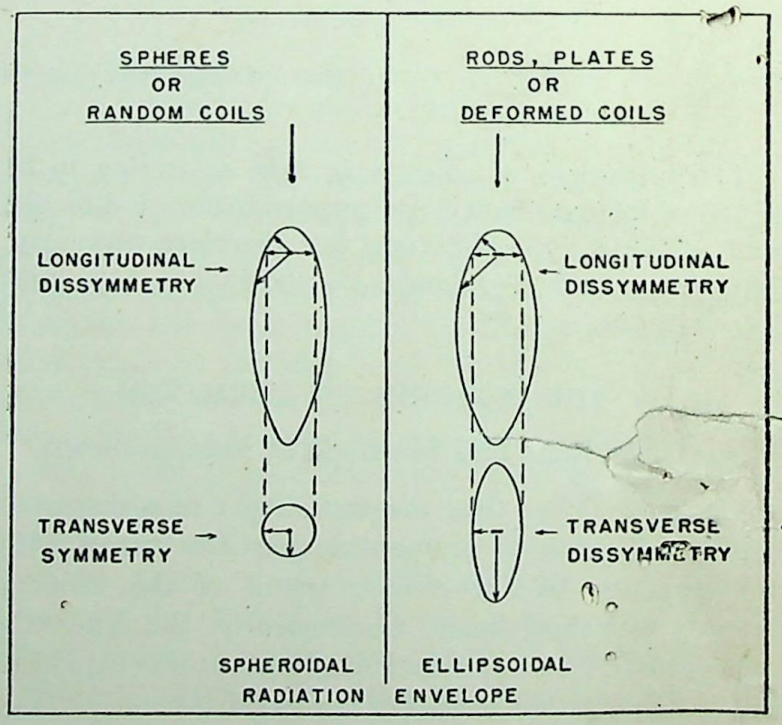


FIG. 5. Conventional dissymmetry (left) and bidissymmetry on orientation or deformation (right).

[24] W. Heller, J. *Phys. Chem.* **41**, 1041 (1937).
[25] W. Heller and G. Quimfe, Compt. rend. **205**, 1152, 1394 (1937); **207**, 157 (1938).
[26] The apparatus constructed and used for this purpose has been described briefly elsewhere[14] and will be discussed in detail in a separate publication.
[27] Details are given in the M. S. *thesis of G. Quimfe*, Faculté des Sciences, University of Paris (1938).
[28] The superficial similarity with rotation due to rotatory power has led a few authors to confuse dichroism with optical rotation. Rotation of the polarizer allows a very easy differentiation between the two effects.
[29] W. Heller and H. Zocher, Z. physik. Chem. (Leipzig) **A164**, 55 (1933).
[30] If interference in the forward direction cannot be neglected—nearly microscopic particle size or large refractive index difference between particles and medium—all three configurations of both rods and disks will lead to different scattering intensities.

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

(right), not negligibly small compared to the wavelength, are given by the upper drawings for the direction of the primary beam indicated by the top arrows. The cross sections of the diagrams in the lower pair of drawings apply to incidence of the primary beam perpendicular to the plane of the paper. The radiation diagram on the left exhibits merely the well-known "longitudinal" dissymmetry, while that on the right has, in addition, a "transverse" dissymmetry. Light scattering is here therefore "bidissymmetrical". The dityndallism in polarized light also exhibits bidissymmetry, but its features are more complicated.

While significant details of the scattering process are lost when unpolarized light is used, the resulting greater simplicity of the phenomena allows simplification of both the optical theory and the apparatus. One may, therefore, consider approximating treatments which would fail for polarized light. Thus, Debye's "interference" treatment[2] was used in order to establish the theory of bidissymmetry expected for flow-deformed random coils.[18] This theory allows one to derive, from the bidissymmetry, important information on molecular parameters which cannot be obtained easily by any other means.

Gurukul Kangri Vishwavidyalaya Library

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

12457

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA

CC-0. Gurukul Kangri University Haridwar Collection. Digitized by S3 Foundation USA